Classification of Intrinsically Disordered Regions and Proteins

Robin van der Lee[1,2,*], Marija Buljan[1,+], Benjamin Lang[1,+], Robert J. Weatheritt[1,+], Gary W. Daughdrill[3], A. Keith Dunker[4], Monika Fuxreiter[5], Julian Gough[6], Joerg Gsponer[7], David T. Jones[8], Philip M. Kim[9,10,11], Richard W. Kriwacki[12], Christopher J. Oldfield[4], Rohit V. Pappu[13], Peter Tompa[14,15], Vladimir N. Uversky[16,17], Peter E. Wright[18] and M. Madan Babu[1,*]

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, United Kingdom
[2]Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Centre, 6500 HB Nijmegen, The Netherlands
[3]Department of Cell Biology, Microbiology, and Molecular Biology, University of South Florida, 3720 Spectrum Blvd., Suite 321, Tampa, Florida, 33612
[4]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA
[5]DE OEC-Momentum Laboratory of Protein Dynamics, Department of Biochemistry and Molecular Biology, University of Debrecen, H-4032 Debrecen, Nagyerdei krt 98, Hungary
[6]Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol BS8 1UB, UK
[7]Department of Biochemistry and Molecular Biology, Centre for High-Throughput Biology, University of British Columbia, Vancouver, BC, Canada
[8]Bioinformatics Group, Department of Computer Science, University College London, London, WC1E 6BT, UK
[9]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada M5S 3E1
[10]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada M5S 3E1
[11]Department of Computer Science, University of Toronto, Toronto, ON, Canada M5S 3E1
[12]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN, USA.
[13]Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130
[14]VIB Department of Structural Biology, Vrije Universiteit Brussel, Brussels, Belgium
[15]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
[16]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[17]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russia
[18]Department of Integrative Structural and Computational Biology and Skaggs Institute of Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

**Corresponding authors: madanm@mrc-lmb.cam.ac.uk and rvdlee@mrc-lmb.cam.ac.uk*

+These authors contributed equally to this work.

Table of contents

Abstract....4
1. Introduction....4
1.1. Uncharacterized protein segments are a source of functional novelty....4
1.2. Structure-function paradigm enhances function prediction....4
1.3. Classification further facilitates function prediction....4
1.4. Intrinsically Disordered Regions and Proteins....5
1.5. The need for classification of Intrinsically Disordered Regions and Proteins....5
2. Function....6
2.1. Entropic chains....6
2.2. Display sites....6
2.3. Chaperones....6
2.4. Effectors....7
2.5. Assemblers....7
2.6. Scavengers....7
3. Functional features....7
3.1. Linear motifs....8
3.2. Molecular recognition features....9
3.3. Intrinsically disordered domains....9
3.4. Continuum of functional features....10
4. Structure....10
4.1. Structural continuum....10
4.2. Protein quartet....11
4.3. Supertertiary structure....11
5. Sequence....11
5.1. Sequence-ensemble relationships....12
5.2. Prediction flavors....12
5.3. Disorder-complexity space....13
5.4. Overall degree of disorder....13
5.5. Length of disordered regions....13
5.6. Position of disordered regions....13
5.7. Tandem repeats....13
6. Protein interactions....14
6.1. Fuzzy complexes....14
6.2. Binding plasticity....15
7. Evolution....15
7.1. Sequence conservation....16
7.2. Lineage and species specificity....17
7.3. History and mechanism of repeat evolution....17
8. Regulation....17
8.1. Expression patterns....17
8.2. Alternative splicing....18
8.3. Degradation kinetics....18
9. Biophysical properties....19
9.1. Phase transition....19
10. Discussion....19
10.1. Current methods for function prediction of IDRs and IDPs....20
10.1.1. Linear motif-based approaches....20
10.1.2. Molecular recognition feature-based approaches....21
10.1.3. Intrinsically disordered domain-based approaches....21
10.1.4. Other approaches....21
10.2. Requirement for annotation....21
10.3. Integration of methods for finding IDR and IDP function....21
10.4. Future directions....23
11. Conclusion....23
Acknowledgements....23
Boxes....25

Box 1: Databases of intrinsically disordered regions and proteins 25
Box 2: Experimental detection of disordered regions 25
Box 3: Evolution of disordered regions 26
Tables 27
Figure legends 32
References 36

Abstract

A large proportion of the protein coding sequences in any genome lack functional characterization. Knowledge about the function of these uncharacterized regions can provide important insights into biological processes. Traditional classifications and methods for predicting protein function are largely based on principles that apply to structured domains that are well folded. These approaches do not explicitly consider Intrinsically Disordered Regions (IDRs) and Intrinsically Disordered Proteins (IDPs), which are polypeptide segments that do not adopt a defined tertiary structure. IDRs and IDPs are ubiquitous in any organism, are prevalent in uncharacterized proteins, and are different from structured domains in many ways. Classification schemes of IDRs are needed to enhance the functional characterization of this important class of protein segments. This review provides an overview of classifications of IDRs and IDPs. We discuss approaches that are based on function, functional elements, structure, sequence, protein interactions, evolution, regulation, and biophysical properties. We conclude with a discussion of resources currently available for gaining insight into IDR function and speculate how the combination of classifications could achieve high quality function prediction for disordered regions in proteins.

1. Introduction

1.1. Uncharacterized protein segments are a source of functional novelty

Over the last decade, we have observed a massive increase in the amount of information describing protein sequences from a variety of organisms.[1,2] While this may reflect the diversity in the sequence space, and possibly also in the function space,[3] a large proportion of the sequences lacks any useful functional annotation.[4,5] Often these sequences are annotated as putative or hypothetical proteins and for the majority their functions still remain unknown.[6,7] Suggestions about potential protein function, primarily molecular function, often come from computational analysis of their sequences. For instance, homology detection allows for the transfer of information on well-characterized protein segments to those with similar sequences that lack annotation of molecular function.[8-10] Other aspects of function, such as the biological processes proteins participate in, may come from genetic- and disease-association studies, expression and interaction network data, and comparative genomics approaches investigating genomic context.[11-17] Characterization of unannotated and uncharacterized protein segments is expected to lead to the discovery of novel functions as well as provide important insights into existing biological processes. In addition, it is likely to shed new light on molecular mechanisms of diseases that are not yet fully understood. Thus, uncharacterized protein segments are likely to be a large source of functional novelty relevant for discovering new biology.

1.2. Structure-function paradigm enhances function prediction

Traditionally, protein function has been viewed as critically dependent on the well-defined and folded three-dimensional structure of the polypeptide chain. This classical structure-function paradigm (**Figure 1**; left panel) has mainly been based on concepts explaining the specificity of enzymes, and protein structure determination methods utilizing X-ray diffraction on protein crystals. The classical concept implies that protein sequence defines structure, which in turn determines function, i.e. function can be inferred from the sequence and its structure. Even when proteins sequences diverge during evolution, for example after gene duplication, the overall fold of their structures remains roughly the same. Therefore, structural similarity between proteins can reveal distant evolutionary relationships that are not easily detectable using sequence-based methods.[18,19] Structural genomics efforts such as the Protein Structure Initiative (PSI) have been set up to enlarge the space of known protein folds and their function, thereby complementing sequence-based methods in an attempt to fill the gap of sequences for which there is no functional annotation.[20,21] Specifically, phase two of the PSI aimed to structurally characterize proteins and protein domains of unknown function, often providing the first hypothesis about their function and serving as a starting point in their experimental characterization.

1.3. Classification further facilitates function prediction

Classification schemes provide a guideline for systematic function assignment to proteins. Generally, proteins are made up of a single or multiple domains that can have distinct molecular functions. These domains, which are referred as structured domains, often fold independently, make precise tertiary contacts and adopt a specific three-dimensional structure to carry out their function. The sequences that compose structured domains can be organized into families of homologous sequences, whose members are likely to share common evolutionary relationship and molecular function. The Pfam database classifies known protein sequences and contains almost 15,000 such families, for most of which there is some understanding about the function.[22] Nevertheless, Pfam also contains more than 3,000 families annotated as domains of unknown function, or DUFs.[23] These families are largely made up of hypothetical proteins and await function annotation. Another powerful example of a protein classification scheme is the Structural Classification of Proteins (SCOP), which provides a means of grouping proteins with known structure together, based on their structural and evolutionary relationships.[24,25] SCOP utilizes a hierarchical classification consisting of four levels, (i) family (ii) superfamily (ii) fold and (iv) class, with each level corresponding to different degrees of structural similarity and evolutionary relatedness between members. Using this scheme, function of newly solved structures or sequences can be inferred from their similarity with existing protein classes through structure or sequence comparisons, for instance as available via the SUPERFAMILY database.[10] In this direction, another major initiative is Genome3D, which is a collaborative project to annotate genomic sequences with predicted 3D structures based on CATH[26] (Class, Architecture, Topology, Homology) and SCOP[24,25] domains to infer protein function.[27]

1.4. Intrinsically Disordered Regions and Proteins

While many proteins need to adopt a well-defined structure to carry out their function, a large fraction of the proteome of any organism consists of protein segments that are not likely to form a defined three-dimensional structure, but are nevertheless functional.[28-41] These polypeptide segments are referred to as intrinsically disordered regions (IDRs; **Figure 1**; right panel).[42] Since IDRs generally lack bulky hydrophobic amino acids, they are unlikely to form the well-organized hydrophobic core that makes up a structured domain[31,43] and hence they follow rules that are different from the classical structure-function view. In this framework, protein sequences in a genome can be viewed as modular because they are made up of combinations of structured and disordered regions (**Figure 1**; bottom panel). Proteins without IDRs are called structured proteins and proteins with entirely disordered sequences that do not adopt any tertiary structure are referred to as intrinsically disordered proteins (IDPs). The majority of eukaryotic proteins are made up of both structured and disordered regions and both are important for the repertoire of functions that a protein can have in a variety of cellular contexts.[42] Traditionally, IDRs were considered to be passive segments in protein sequences that 'linked' structured domains. However, it is now well established that IDRs actively participate in diverse functions mediated by proteins. For instance, disordered regions are frequently subjected to post-translational modifications (PTMs) that increase the functional states in which a protein can exist in the cell.[44,45] In addition, they expose short linear peptide motifs of about 3-10 amino acids that permit interaction with structured domains in other proteins.[46,47] These two features in isolation or in combination permit interaction and recruitment of diverse proteins in space and time, thereby facilitating regulation of virtually all cellular processes.[46] The prevalence of IDRs in any genome (see for example the D^2P^2 database,[48] **Box 1**) in combination with their unique characteristics means that these regions extend the classical view of the structure-function paradigm and hence that of protein function. Thus, functional regions in proteins can either be structured or disordered, and these need to be considered as two fundamental classes of functional building blocks of proteins.[49]

1.5. The need for classification of Intrinsically Disordered Regions and Proteins

IDRs and IDPs are prevalent in eukaryotic genomes. For instance, 44% of human protein-coding genes contain substantial disordered segments of >30 amino acids in length (**Figure 2A**).[48] In the human genome, 6.4% of all protein-coding genes do not have any function annotation in their description in ENSEMBL[1] (**Figure 2B**). Further investigation using the D^2P^2 database of disorder in genomes[48] revealed that most of these genes with no function annotation encode at least some disorder (**Figure 2B**) and that genes with no annotation contain proportionally more IDRs (**Figure 2C**). Given the absence of structural constraints, IDRs tend to evolve more rapidly than protein domains that adopt defined structures.[50-55] As a result, identifying homologous regions is harder for IDRs and IDPs than it is for structured domains. This complicates the transfer of function information between homologs and the prediction of function of IDRs and IDPs. Furthermore, much of protein annotation is based on information on sequence families and structured domains. However, less than half of all residues in the human proteome fall within such domains (**Figure 3**). Not only do most residues of human proteins fall outside domains, a large fraction of these residues are also disordered (**Figure 3A** and **3B**, right bars). Furthermore, although it is expected that SUPERFAMILY domains based on known protein structures have very little disorder (**Figure 3A**, left bar), Pfam domains based on sequence clustering do not contain much more (**Figure 3B**, left bar). These observations suggest that there is a large pool of protein segments that are not considered by conventional protein annotation methods, because the sequences of disordered regions are difficult to align, or because the methods do not explicitly consider disordered and non-domain sequences. Taken together, these considerations raise the need to devise a classification scheme specifically for disordered regions in proteins that may enhance the function prediction and annotation for this important class of protein segments.

In this review, we synthesize and provide an overview of the various classifications of Intrinsically Disordered Regions and Proteins that have been put forward in the literature since the start of systematic studies into their function some 15 years ago. We discuss approaches based on function, functional elements, structure, sequence, protein interactions, evolution, regulation, and biophysical properties (**Table 1**). Finally, we discuss resources that are currently available to gain insight into IDR function, we suggest areas where increased efforts are likely to advance our understanding of the functions of protein disorder, and we speculate how combinations of multiple existing classifications could achieve high quality function prediction for IDRs, which ultimately leads to improved function coverage and a deeper understanding of protein function.

2. Function

Dunker and co-workers[56] distinguished 28 separate functions for disordered regions, based on literature analysis of 150 proteins containing disordered regions of 30 residues or longer. These functionalities can be summarized as molecular recognition, molecular assembly, protein modification, and entropic chains. Further development of this scheme resulted in one comprising six different function classes of disordered protein regions: entropic chains, display sites, chaperones, effectors, assemblers, and scavengers (**Figure 4**).[33,57] In line with this classification, Gsponer and Babu classified IDR function into three broad functional categories: (i) facilitated regulation via diverse post-translational modifications, (ii) scaffolding and recruitment of different binding partners, and (iii) conformational variability and adaptability (**Figure 5**).[38] A single protein may consist of several disordered regions that belong to different function classes, which may be associated with a specific position in the sequence.[58] The following section will address and exemplify the six functionalities of disordered regions.

2.1. Entropic chains

Entropic chains carry out functions that benefit directly from their conformational disorder, i.e. they function without ever becoming structured. Examples of entropic chains include flexible linkers, which allow movement of two domains relative to each other, and spacers that regulate the distances between domains. Evidence that flexibility is a functional characteristic that needs to be maintained came from studies on a family of flexible linkers in the 70 kDa subunit of replication protein A (RPA70), which display conserved dynamic behavior in the face of negligible sequence conservation.[59] The microtubule-associated protein 2 (MAP2) projection domain exemplifies spacer behavior as it repels molecules that approach microtubules, thereby providing spacing in the cytoskeleton. Another subcategory of entropic chains are entropic springs, such as those present in the titin protein, which contains repeat regions rich in PEVK amino acids that generate force upon over-stretching to help restore muscle cells to their relaxed length.[60,61]

2.2. Display sites

Post-translational modifications (PTMs) affect the stability, turnover, interaction potential and localization of proteins within the cell.[62] These aspects of PTMs are particularly relevant for proteins involved in regulation and signaling, as are many IDPs.[35,36,38,63,64] The conformational flexibility of disordered protein regions as display sites provides obvious advantages over structured regions. Flexibility facilitates (i) the deposition of PTMs by enabling transient but specific interaction with catalytic sites of modifying enzymes, and (ii) easy access and recognition by PTM-binding proteins that mediate downstream outcomes upon binding.[46,65] Indeed, experimental and computational approaches have shown that disordered regions are enriched for sites that can be phosphorylated.[44,45,66] Results of computational studies suggest that IDPs are likely to be substrates of a large number of kinases and other modifying enzymes as they are heavily post-translationally modified.[45,67,68] Furthermore, PTM sites are often located within short peptide motifs, modification of which influences the affinity for interaction with diverse binding partners (see **section 3.1.**).[69,70] In turn, disordered protein regions are strongly enriched for these motifs,[46,71-73] underlining the importance of intrinsic disorder as PTM display sites. Well-characterized examples of IDPs in which PTMs are key to function and regulation include, among others, histones, p53, and the cyclin-dependent kinase regulator p27.[74-76]

2.3. Chaperones

Chaperones are proteins that assist RNA and protein molecules to reach their functionally folded states.[77,78] Statistical analysis showed that disordered regions make up over half of RNA chaperones and over a third of protein chaperones.[79,80] The versatility of disordered segments seems extremely suitable for chaperone mechanics, although mechanistic evidence is still scarce.[81] Firstly, their capacity to adapt to many different binding partners matches the need for chaperones to bind a wide range of proteins. Secondly, disordered segments enable quick interactions, for example preventing misfolded proteins from forming toxic aggregates by binding them quickly. Finally, the binding thermodynamics of disordered regions are well suited for the cycles of repeated chaperone binding and release that enable substrate folding. It has been proposed that binding of disordered chaperone regions to misfolded substrates induces local folding of the disordered chaperone, and promotes unfolding of the substrate, thereby providing the substrate with a chance to refold correctly. This reversible exchange of entropy between chaperone IDRs and misfolded substrates forms the basis of the entropy transfer model[79] and represents a distinct type of chaperone function that relies on disordered regions and does not require ATP. This mechanism can even be switched on and off at need by regulated transitions between folded and disordered states, as reported in the case of the redox-regulated chaperone Hsp33.[82]

2.4. Effectors

Another functional class of disordered regions is that of the effectors, which interact with other proteins and modify their activity. Upon binding their interaction partners, IDRs often show a disorder-to-order transition, also known as coupled folding and binding.[83,84] Examples of two effectors that fold upon binding are p21 and p27, which regulate different cyclin-dependent kinases (Cdk) that are responsible for the control of cell-cycle progression in mammals.[65] p21 and p27 exhibit functional diversity by achieving opposite effects on different Cdk-cyclin complexes; promoting the assembly and catalytic activity of some (e.g. Cdk4 paired with D-type cyclins), and inhibiting others (e.g. Cdk2 paired with A- and E-type cyclins).[65] Another effector IDP is calpastatin, which undergoes significant folding upon binding calpain, thereby achieving specific and reversible inhibition.[85] Effector IDRs can also affect the activity of other parts within the same protein for example to achieve auto-inhibition. The intrinsically disordered GTPase-binding domain (GBD) of the Wiskott–Aldrich syndrome protein (WASP) illustrates such behavior.[86] Binding of the GBD to the Cdc42 protein promotes interaction of WASP with the actin cytoskeleton regulatory machinery. However, GDB adopts a different structure when it folds back on other parts of WASP to inhibit actin interaction. More generally, autoinhibitory regions are indeed enriched for intrinsic disorder and often have different structures in the inhibitory and active states of the protein.[87]

2.5. Assemblers

Disordered assemblers bring together multiple binding partners to promote the formation of higher-order protein complexes, such as the ribosome (many ribosomal proteins are disordered[88]), activated T-cell receptor complexes and the transcription preinitiation complex.[57,89] The presence of multiple molecular recognition features (MoRFs) and short linear peptide motifs (SLiMs) within the disordered segments enables binding and can bring together different partners (see **section 3.1.** and **3.2.**). Indeed, larger complexes are assembled from proteins that tend to be more disordered[90] and intrinsic disorder is a common feature of hubs in protein interaction networks.[91,92] The open structure of disordered assemblers is largely preserved upon scaffolding their partner proteins, resulting in a large binding interface that enables multiple proteins to be bound by a single IDR.[93] Furthermore, intrinsic disorder avoids the problem of steric hindrance that prevents the formation of comparably large complexes from structured proteins. Assembler function can be imagined in two ways: (i) as structural mortar that helps bring together proteins by stabilizing the complexes they form, and (ii) as a scaffold that serves as backbone for the spatio-temporally regulated assembly of different signaling partners, allowing mediation of signaling events as initiated by specific stimuli and external conditions. An example of the latter mechanism is the axis inhibition (Axin) scaffold protein, which colocalizes β-catenin, casein kinase Iα, and glycogen synthetase kinase 3β by their binding to Axin's long intrinsically disordered region, thereby effectively yielding a complex of structured domains with flexible linkers.[94] The assembly of all four proteins accelerates interactions between them by raising their local concentrations and leads to the efficient phosphorylation and subsequent destruction of β-catenin. Scaffolding regions have one of the highest degrees of disorder of all function categories.[95,96]

2.6. Scavengers

The final distinct function class of IDRs and IDPs are scavengers, which store and neutralize small ligands. Chromogranin A, one of the earliest examples of an IDP, functions as a scavenger by storing ATP and adrenaline in the medulla of the adrenal gland.[97] NMR studies showed that chromogranin is a random coil in both the isolated form and in its cellular environment in the intact adrenal gland.[97] Other examples of scavenger IDPs are casein, a protein that binds clusters of calcium phosphate, and salivary proline-rich glycoproteins, which bind tannin molecules in the digestive tract.[33]

3. Functional features

Different types of functional regions in intrinsically disordered proteins have been uncovered both by investigations aimed directly at increasing the understanding of IDRs, and indirectly by linking previously studied functionality of proteins to disordered regions. First, the majority of linear motifs (such as the SH2 domain interaction motif) have been found as enriched in IDRs.[47,71,98] Second, the development of disorder prediction methods has led to the identification of segments that promote disorder-to-order transitions,[99-103] which have been verified using known crystal structures. Third, some interaction domains identified using crystallography, by sequence analysis, and by other techniques, turn out to be intrinsically disordered in solution (e.g. the BH3 domain[104]). The following section discusses these three interaction features separately and points out the underlying connections between them.

3.1. Linear motifs

A common functional module within IDRs is the linear motif,[46,47,71] also known as LMs, short linear motifs (SLiMs),[105] or MiniMotifs.[106] By regulating low-affinity interactions, these sequence motifs (annotated instances are usually 3-10 amino acids long[47]) target proteins to a particular subcellular location, recruit enzymes that alter the chemical state of the motif by post-translational modifications (PTMs), control the stability of a protein, and promote recruitment of binding factors to facilitate complex formation.[46,47] Linear motifs, helped by the flexible nature of the disordered regions that surround them,[71] primarily bind onto the surfaces of globular domains[107,108] and their compact binding surface promotes them to occur multiple times within one protein.[46,47] Moreover, the short nature of many linear motifs means they have a high propensity to convergently evolve and emerge in unrelated proteins.[46,47] A consequence of these properties is that pathogenic viruses and bacteria have evolved to mimic these linear motifs, allowing them to manipulate regulation of cellular processes.[109,110]

Linear motifs can be broadly divided into two major families: those that act as modification sites and those that act as ligands, with each having numerous subgroups (Van Roey et al. *Chemical Reviews*, same issue) (**Figure 6**). The first major family, the enzyme binding or modification motifs, can be divided into three groups. (i) Post-translational processing events or proteolytic cleavage. A well-known example is the motif recognized by Caspase-3 and -7, which has a [ED]xxD[AGS] consensus sequence. Caspases are a family of proteases that promote apoptosis and inflammation by cleaving such motifs in their substrate proteins.[111] Hundreds of proteins have convergently evolved the Caspase-3/-7 motif to become under the regulation of the apoptotic pathway.[112] (ii) PTM moiety removal and addition. Post-translational modifications are often catalyzed by enzymes, many of which recognize a specific binding sequence. For example, the cyclin-dependent kinase recognition motif, [ST]Px[KR], is present in many mitotic proteins and its phosphorylation is key for regulating cell cycle progression.[113] (iii) Structural modifications. This group of motifs is involved in the catalyzed conformational alteration of a peptide backbone. The classic example is the peptidylprolyl cis-trans isomerase (PPIase) Pin1, which binds [ST]P motifs in a phosphorylation dependent manner to catalyze the cis-trans isomerization of the proline peptide bond. This modification can regulate the recognition of [ST]P phosphorylated sites by phosphatases.[114]

The second major family of motifs are the ligand motifs, which can also be divided into three main groups (**Figure 6**). (i) Complex promoting motifs are the most well-known class of motifs and include the phosphorylated tyrosine motif recognized by SH2 (Src homology 2) domains, the C-terminal motifs that bind PDZ domains, and the proline-rich PxxP motifs that interact with SH3 (Src homology 3) domains.[115] These motifs often function in protein scaffolding and their multivalency (tendency to occur multiple times in one sequence) can increase the avidity of interactions promoting phase transition (see **section 9.1.**).[116] (ii) Docking motifs increase the specificity and efficiency of modification events (e.g. addition or removal of PTMs, see above) by providing additional binding surface. These motifs are distinct from the modification sites and are usually (but not always[117]) in the same protein. Examples include the degron motifs (KEN box, D box, F box), which act as recognition surfaces for ubiquitin ligases and often function to regulate protein degradation by the 26S proteasome.[117-119] The KEN box motif occurs in several key mitotic kinases to ensure their degradation or deactivation at mitotic exit.[119] (iii) Targeting motifs can localize proteins towards subcellular organelles. For example, importin proteins of the nuclear pore complex recognize the nuclear localization signal (NLS), usually a motif containing a short cluster of lysines and arginines, and translocate NLS-containing proteins into the nucleus.[120] Targeting motifs can also act to traffic proteins, as in the case of motifs recognized by adaptor proteins at different stages of endocytosis, which ensure that vesicles are trafficked to the right location.[121]

An important feature of linear motifs is their propensity to act as molecular switches. This is for two major reasons: (i) linear motif-mediated interactions are low affinity, which means that large, bulky post-translational modifications will alter their the binding properties;[70] (ii) their small footprint (or surface area) allows motifs to occur multiple times in one sequence promoting high avidity interactions and the recruitment of multiple factors (e.g. the LAT complex in T-cell receptor signaling[122]). This also means two motifs can overlap resulting in mutually exclusive binding.[72] The ability of a motif to rapidly switch between binding partners and create multivalent complexes is crucial for the creation of dynamic signaling networks.[70]

3.2. Molecular recognition features

Disordered segments can also contain another type of peptide motif (10–70 amino acids) that promotes specific protein–protein interactions. These functional elements are called preformed structural elements (PSEs),[99] molecular recognition features (MoRFs) or elements (MoREs),[100-102] or prestructured motifs (PreSMos).[103] Importantly, MoRFs undergo disorder-to-order transition upon binding to their interaction partners (i.e. folding upon binding),[37,101,103] and often the unbound form of these preformed elements is biased towards the conformation that they have in complex.[99] Preformed structural elements and MoRFs may serve as initial contact points for interaction events, which have different kinetic and thermodynamic properties than interactions between structured protein regions. Binding of preformed elements is one version of conformational selection (see **section 6.**), suggested long ago for interactions with flexible ligands.[123] At the other extreme is coupled binding and folding, or induced fit, in which structure formation and binding occur concomitantly after the formation of the initial encounter complex. Given the complexity of many complexes involving intrinsically disordered partners, interactions involving both conformational selection of preformed elements and induced fit likely occur.[84,124]

MoRFs occurring in the Protein Data Bank[125] can be classified into subtypes according to the structures they adopt in the bound state: α-MoRFs, β-MoRFs, and ι-MoRFs (**Figure 7A-C**),[101] which form α-helices, β-strands, and irregular secondary structure when bound, respectively. MoRFs that contain combinations of different types of secondary structure are called complex (**Figure 7D**).[101] The p53 protein contains multiple MoRFs that are disordered in the absence of their interactors (**Figure 7E**).[100,101] The first p53 MoRF is located near the N-terminus and undergoes a transition from a disordered to an α-helical state upon interaction with the Mdm2 protein. In fact, this region of p53 exemplifies the high potential of IDRs for multiple partner binding as it is known to bind more than 40 different partners, although for most of these complexes the 3D structures are not determined and therefore the MoRF type is not always known. The region between p53 residues 40 and 60 features an α-MoRF that functions as a secondary binding site for Mdm2 as well as a primary binding site for RPA70.[126] In the absence of any binding partner, this region shows evidence of minimal helical secondary structure,[127] whereas bound to either Mdm2[128] or RPA70[129] stronger helical structures are formed. The C-terminal region of p53 also contains a MoRF that interacts with multiple partners, giving rise to different bound structures. For example, the S100B(ββ) protein induces a helical structure, while interaction with the Cdk2-cyclin A complex leads to an irregular ι-MoRF. An example of the role of MoRFs in scaffolding proteins is RNase E, which assembles the RNA degradosome.[130] The flexible C-terminal end of RNase E contains several recognition motifs that are central to its scaffolding function and serve as binding sites for other members of the degradosome.[131] For example, an α-MoRF interacts with enolase[132] and a β-MoRF binds polynucleotide phosphorylase.[133] The recognition features are connected by disordered segments that accommodate assembly of the complex by providing flexibility. Lee and co-workers[103] have annotated the secondary structure propensities of many other regions that display transient structural elements and undergo disorder-to-order transitions, all of which have been experimentally confirmed by NMR spectroscopy.

Sequence context can play an active role in modulating the degree of structural pre-organization of a MoRF. An example pertains to the study of DNA binding motifs in the basic regions (bRs) of basic region leucine zipper transcription factors.[134] The bRs are 28-30 residues long regions predicted to be highly disordered and include a strongly conserved ten-residue DNA binding motif (DBM). The α-helicity (i.e. preference for α-helical conformation) of the DBM in the unbound form is modulated by the sequence of the N-terminal segment that is directly in *cis* to the DBM.[134] For example, the N-terminal sequence contexts of Gcn4 and Cys3 DBMs contribute to a higher level of helicity of the DBM than the same region in c-Fos and Fra1 (whose DBMs have a low helicity). Essentially, the N-terminal sequence contexts are helix caps and these can be used in different ways to ensure different levels of structural pre-organization within an α-MoRF, thereby suggesting that sequence contexts can provide useful clues when classifying MoRFs and linear motifs.[135]

3.3. Intrinsically disordered domains

Most protein domains that are identified using sequence-based approaches are structured, but some can be fully or largely disordered[136] or contain conserved disordered regions[137] known as intrinsically disordered domains (IDDs). For instance, about 14% of Pfam domains have more than 50% of their residues in predicted disorder, and many well-known domains, such as the kinase-inhibitory domain (KID) of Cdk inhibitors (e.g. p27[65]) and the Wiskott–Aldrich syndrome protein (WASP)-homology domain 2 (WH2) of actin-binding proteins[136] have been shown experimentally to be fully disordered. Protein domains with conserved disordered regions have a variety of functions, but are most commonly involved in DNA-, RNA- and protein binding.[137] Furthermore, domains that were gained during evolution by the extension of existing exons contain the highest degree of disordered regions.[138] This suggests that exonization of previously non-coding regions could be an important mechanism for the addition of disordered segments to proteins.

Interestingly, it has also been observed that particular disordered regions frequently co-occur in the same sequence with specific protein domains.[139,140] Some domain families appear only to require the presence of disorder in their neighborhood for functioning, while others seem to rely on the occurrence of disordered regions in specific locations relative to the start or end of the protein domain.[139] For example, particular combinations of domains, involved mainly in regulatory, binding, receptor, and ion-channel roles, only occur with a disordered region inserted between them, while others only occur without a disordered domain between them. These observations imply that short disordered regions in the vicinity of protein domains complement the function of a structured domain, and in some cases may comprise separate functional domains in their own right. Thus, the co-occurrence of IDRs and structured domains in the same protein might be useful to gain insight into unannotated disordered regions.

3.4. Continuum of functional features

A measure that is often used to distinguish the different types of disordered binding modules is length, however this is likely to stem primarily from the different methodology used for their detection. Protein domain detection relies on hidden Markov models,[22] which are poor at identifying short sequences, and therefore domain annotation tends to focus on larger binding modules. In contrast, linear motifs in the ELM database[105] (based on which average linear motif length is defined) are biased towards short binding modules as these more straightforward to annotate. Finally, the tendency of MoRFs and preformed elements to undergo disorder-to-order transition and the statistics used for their detection, means that these features tend to be slightly longer than annotated linear motifs.

Thus, although there are differences in the definitions of linear motifs and MoRFs, they share many common features[71,141] including a tendency to undergo disorder-to-order transition (all MoRFs by definition and 60% of LMs[47]), an enrichment in IDRs (MoRFs by definition and 80% of LMs are in IDRs[47,71]), and a tendency to promote complex formation.[47,102] Intrinsically disordered domains can also have significant overlap with MoRFs and linear motifs. For example, the WH2 domain is considered an IDD[136] and is also defined as a motif in the ELM database.[105] One feature that is probably unique to IDDs is that some are not only capable of binding to well-folded, structured domains (a mechanism shared with motifs and MoRFs), but can also bind each other in a process of mutually induced folding. For example, the nuclear coactivator binding domain (NCBD) of CREB-binding protein (CBP) and the activator for thyroid hormone and retinoid receptors (ACTR) domain of p160 are both disordered on their own but upon interaction form a complex by mutual synergistic folding.[142] The overlap between especially linear motifs and MoRFs, but also IDDs, suggests that these functional features are different states in the same continuum of binding mechanisms involving disordered regions.

4. Structure

Intrinsically disordered regions and proteins show a wide variety of structural subtypes. These different types of disorder can be characterized using an array of experimental techniques (**Box 2**) and several resources collect computationally identified and experimentally verified disordered regions (**Box 1**). The following section discusses classification schemes that are based on structural features of disordered proteins.

4.1. Structural continuum

Proteins have been proposed to function within a conformational continuum, ranging from fully structured to completely disordered.[36] The spectrum covers tightly folded domains that display either no disorder or only local disorder in long loops or tails, multidomain proteins linked by disordered regions, compact molten globules containing extensive secondary structure, collapsed globules formed by polar sequence tracts, unfolded states that transiently populate local elements of secondary structure, and highly extended states that resemble statistical coils (**Figure 8**). In this model, there are no boundaries between the described states and native proteins could appear anywhere within the continuous landscape. IDRs are highly dynamic and fluctuate rapidly over an ensemble of heterogeneous conformations.[143] Thus, an IDR may fluctuate stochastically between several different states, transiently sampling coil-like states, localized secondary structure, and more compact globular states. Transient localized elements of secondary structure (most often helix) are common in amphipathic regions of the sequence and this potentially plays a role in binding processes.[84] The structural characteristics and populations of the individual states in the conformational ensemble and the degree of compaction of the polypeptide chain are determined by the nature of the amino acids and their distribution in the IDR sequence (see **section 5.1.**).[144,145]

4.2. Protein quartet

The protein quartet model proposed by Uversky suggests that protein function can arise from four types of conformational states and the transitions between them (**Figure 9**). The conformational states can be classified as random coil, pre-molten globule, molten globule, and folded.[34] In this model, unbound disordered regions could fall into all categories except for 'folded'. Originally it was shown that IDPs can realize extended conformations (i.e. random coil-like) or remain globally collapsed (i.e. molten globule-like) with regions of fluctuating secondary structure.[32] Hydrodynamic radii and far-UV circular dichroism (CD) experiments revealed that random coil-like IDPs can be further subdivided into intrinsic pre-molten globules and intrinsic coils.[34] Proteins in the pre-molten globule state are less compact than molten globules, but still show some residual secondary structure. In contrast, proteins in the intrinsic coil state show little or no secondary structure. The average net charge in part determines the conformational state of a disordered region, with low and high average charges corresponding to disordered globules and swollen coils, respectively (see **section 5.1.**).[144,145] Furthermore, the pre-molten globule state has a high propensity to participate in folding upon binding events,[146] which would make this structural state suitable for disordered regions acting as effectors and scaffolds. Thus, IDRs are able to undergo structural transitions, for example upon binding to partner molecules, with conformational rearrangements varying across the whole range of conformational states. The quartet model was further extended based on the notion that IDPs and IDRs possess remarkable structural and sequence heterogeneity. IDRs may be considered as modular assemblies of foldons (independently foldable regions), inducible foldons (foldable regions that can gain structure as a result of interaction with specific partners), semi-foldons (regions that are always partially folded), and non-foldons (regions that that do not fold at all).[147] In other words, as described previously in the structural continuum model (see **section 4.1.**),[36] there is a continuous spectrum of differently disordered conformations extending from fully ordered to completely structure-less proteins, with everything in between.

FG nucleoporins are an example of the functional significance that different disordered conformations can have. The porins make up the central part of Nuclear Pore Complexes (NPCs) and regulate nucleocytoplasmic transport.[148] Intrinsically disordered regions with multiple phenylalanine-glycine (FG) motifs make up large parts of the NPC gates. FG domains adopt various disordered conformations with specific functions.[149] Some domains have the low charge characteristics of collapsed coils, while others are characterized by a high degree of charged amino acids, giving rise to relaxed and extended coil structures. Molecular dynamics simulations have shown that extended coils are more dynamic than collapsed coils, suggesting distinct functionalities for the two structural groups. Interestingly, some FG nucleoporins feature both types of disorder along their polypeptide chain. Combinations of disorder subtypes in nucleoporin domains are likely to contribute to NPC gating behavior by creating 'traffic' zones with distinct physicochemical properties that influence the dynamics of substrate translocation through the nuclear envelope.[149-152]

4.3. Supertertiary structure

IDRs allow for complex regulatory phenomena, as witnessed in the case of multidomain proteins in signaling and regulation. Due to the presence of structural disorder, functional domains and short motifs, multidomain proteins are characterized by a dynamic ensemble of tertiary conformations. Some conformations are dominated by intra-molecular domain-domain and domain-motif interactions and are closed and structured in nature, while other conformations are more open and disordered. This state of intra-molecular and conformational variability lays in between the tertiary structure of domains and quaternary structure of multiprotein assemblies, and has been termed supertertiary structure.[153] Complex regulatory function stems from transitions in the ensemble of these structures, as witnessed in the case of several well characterized proteins, such as the Wiskott–Aldrich syndrome protein (WASP, see **section 2.4.**),[86] the Src-family tyrosine kinase Hck,[154] and the E3 ubiquitin ligase Smurf2.[155]

5. Sequence

The sequences of IDPs and IDRs have distinct compositional biases such as the enrichment in charged and polar amino acids and deficiency of bulky hydrophobic groups.[31,43,156,157] These biases have lead to the inference that disorder is a natural consequence of weakening the hydrophobic forces that drive folding of polypeptides into compact tertiary structures. Although disordered regions generally lack the ability to fold independently due to these biases in amino acid composition, within IDRs distinct subsets of sequences can be identified that have different structural and functional characteristics. The following section covers sequence-based classification schemes of IDRs.

5.1. Sequence-ensemble relationships

Systematic efforts combining experiments and computations have addressed the relationship between information encoded in amino acid sequences and the ensemble of conformations these sequences can sample in different conditions. These studies have focused on three major archetype sequences, namely polar tracts, polyelectrolytes, and polyampholytes.[158] Polar tracts are sequence stretches enriched in polar amino acids such as glutamine, asparagine, serine, glycine, and proline and deficient in charged as well as hydrophobic residues. These polar tracts (especially glutamine, asparagine, and glycine-rich sequences) form globules that are generally devoid of significant secondary structure preferences[159-162] and can be as compact as well-folded domains.[158] Collapse of polar tracts arises from the preference for self-solvation over solvation by the aqueous milieu and disorder derives from a lack of specificity for a single compact conformation as instead heterogeneous ensembles of conformations with similar stabilities and compactness are formed. In terms of the energy landscape,[163] the free energy surface for polar tracts is weakly funneled and resembles an 'egg carton'. Interestingly, the drive to collapse, which implies a drive to minimize the interface between the IDR and the surrounding solvent, can also give rise to the significant aggregation and solubility problems[164] as is the case with several glutamine, asparagine, and glycine-rich sequences that are implicated in amyloid formation and phase separation.[165]

Another end of the compositional spectrum are polyelectrolytes. Their amino acid compositions are biased toward charged residues of one type such as the arginine-rich protamines[144] or the Glu/Asp-rich prothymosin α.[145] Experiments and simulations have shown that the tendency of polypeptide backbones to form ensembles of collapsed structures can be reversed by increasing the net charge per residue past a certain threshold (**Figure 10A**). The transition between globules and expanded coils is sharp, suggesting that small changes to the net charge per residue through post-translational modifications such as serine or threonine phosphorylation or lysine acetylation can cause reversible globule-to-coil transitions. These transitions might control the accessibility of SLiMs and MoRFs or even modulate the conformations of these elements.

The impact of the net charge per residue on the conformational properties of IDPs can be summarized in a diagram-of-states (**Figure 10A**),[144] which generalizes the original charge-hydropathy plot.[31] The diagram classifies IDPs based on their amino acid compositions. Annotation using curated disordered sequences from the DisProt database[166] (**Box 1**) suggests that a vast majority (~95%) of IDPs have amino acid compositions that predispose them to be globule formers (**Figure 10A**).[167] This inference would imply that the majority of IDPs are unstable proteins that pose serious challenges for protein homeostasis, because globule-forming sequences do not fold autonomously and should be marginally soluble. Importantly however, a majority (~75%) of the predicted globule formers are actually polyampholytes in that they are enriched in charged residues but have roughly equal numbers of positive and negative charges. Although such sequences are classified as globule formers on the basis of their low net charge per residue, in reality the conformational properties of polyampholytes are governed by the linear sequence distribution of oppositely charged residues. If the oppositely charged residues are segregated in the linear sequence, then electrostatic attractions between oppositely charged blocks cause chain collapse and result in hairpin or globular conformations. In sequences with well-mixed oppositely charged residues, the effects of electrostatic repulsions and attractions counterbalance. These mixed sequences adopt random-coil or globular conformations, depending on the total charge (in terms of the fraction of charged residues) (**Figure 10B**). Many IDPs are strong polyampholytes with well-mixed linear patterns of oppositely charged residues.[167] Thus, IDPs are actually enriched in different classes of random coils that form swollen, loosely packed conformations (**Figure 10B**). Such random-coil sequences are likely to help improve the solubility profiles of connected structured domains and to promote the flexibility that is required for functions like entropic tethers or bristles. These biophysical principles of sequence-ensemble relationships enable the use of de novo sequence design as a tool for modulating these relationships and assessing their impact on functions associated with IDPs and IDRs.

5.2. Prediction flavors

Methods to predict disorder from sequence can be classified into two categories:[168,169] (i) methods that rely on the physical principles underlying protein folding and (ii) methods that are based on machine learning approaches (e.g. support vector machines or neural networks) and are trained on available experimental data which indicates the absence of structure in proteins (for instance, the lack of electron density in crystallographic data). Disorder prediction accuracy varies for different types of disordered regions.[170] Some predictors accurately predict certain disordered regions but have low accuracy predicting others, whereas other predictors give opposite results. Vucetic and co-workers[170] classified protein disorder into three different 'flavors' based on competition between disorder predictors. These V, C, and S disorder flavors (corresponding to the names of the disorder predictors that best predict them: VL-2V, VL-2C, and VL-2S) show differences in sequence composition and combinations of flavors could be associated with different protein functions. For example, disordered regions that bind to other proteins are enriched for flavor S, while disordered ribosomal proteins belong to flavor V. Flavor C gave strong disorder predictions for sugar binding domains.

5.3. Disorder-complexity space

The relationship between sequence complexity and disorder propensity provides further insight into the structural and function variations of IDRs.[171] Different function classes of proteins often show a different disorder-complexity (DC) space distribution. A frequently observed DC-trace is composed of a compact structured region and a section extending out into low complexity, disordered space before looping back into the structured region. This pattern describes a disordered region acting as a linker between structured domains. An example is the bacterial translation initiation factor, which has a disordered linker that appears in the DC-trace as a low-complexity, disordered loop connecting the N- and C-terminal structured, high-complexity domains.[171,172] Thus, functionally related proteins have similar disorder-complexity distributions, suggesting that these distributions might be useful for predicting the function of a disordered region.

5.4. Overall degree of disorder

Large-scale studies into IDP function often group the proteins based on some measure of disorder. For example, protein sequences have been categorized based on the overall degree of disorder or fraction of residues that are shown or predicted to be disordered,[67,173] resulting in e.g. groups of structured proteins (0-10% disorder), moderately disordered proteins (10-30% disorder), and highly disordered proteins (30-100% disorder). Other studies classified proteins based on an overall score of disorder for the whole protein,[174] and the presence or absence of continuous stretches of disordered residues with a specific length.[35,50,139,173] Largely structured proteins are enriched for e.g. metabolic functions, while highly disordered proteins function predominantly in regulation. Hence, classification of disordered proteins based on the level of disorder provides clues about what types of functions are likely.

5.5. Length of disordered regions

The length of IDRs in human follows a power law distribution: there are large numbers of short disordered regions and increasingly smaller numbers of longer ones.[175] Other eukaryotic and prokaryotic proteomes show similar disorder length profiles. Short IDRs may function as linkers and contain individual linear motifs or MoRFs, whereas longer disordered regions might be entropic chains or contain combinations of motifs or domains functioning in recognition. Long disordered regions (more than 500 residues) are typically over-represented in transcription-related functions,[176] whereas proteins containing IDRs of intermediate length (300-500 residues) are enriched for kinase and phosphatase functions. Short IDRs (less than 50 residues) tend to be linked to metal ion binding, ion channel, and GTPase regulatory functions. Thus, the length of a disordered region can also provide a useful indication about the function of the protein containing it.

5.6. Position of disordered regions

Almost all human proteins have some disordered residues within their terminal regions.[58] For example 97% of proteins have predicted disorder in the first or last five residues.[139] Disordered N-terminal tails are common in DNA-binding proteins, and have been shown to contribute to efficient DNA scanning.[177] Furthermore, proteins that are relatively rich in disordered residues at the C-terminus are often associated with transcription factor repressor and activator activities compared to proteins rich in internal or N-terminal disorder.[176] Ion channel proteins in turn are enriched for disordered residues at the N-terminus and the same is true to lesser extent for C-terminal disorder.[176] Indeed, voltage-gated potassium channels often contain disordered residues at their termini, some of which are responsible for channel inactivation.[178] Finally, proteins that are relatively rich in internal disordered regions are weakly enriched for transcription regulator and DNA binding activity.[176] Thus, the relative position of a disordered region in a sequence provides clues about the function of the protein containing it.

5.7. Tandem repeats

Short tandem repeats are common in IDRs and IDPs.[60,179] A study comparing the overlap between amino acid repeats and disordered regions demonstrated that most tandem repeats are associated with disordered regions.[180] For instance, as much as 96% of polyglutamate and polyserine stretches lie within IDRs. Similarly large fractions were found for proline, glycine, glutamine, lysine, aspartate, arginine, histidine, and threonine repeats. In contrast, polyleucine stretches occur predominantly within structured regions. These observations agree with the compositional bias of disordered regions; the most common tandem repeats in IDRs are made up of disorder-promoting residues[43,156] and of sequence patterns that are typically associated with disorder.[157] Moreover, a distinction between perfect and imperfect tandem repeats suggests that as the repeat perfection increases, so does the disorder content.[179]

Repeats of different composition have been linked to specific functions. Consequently, the presence of particular types of repeats is likely to contribute to IDR functioning. Descriptions and examples of different classes of disordered tandem repeats and their structural characteristics have been published previously.[181] For instance, polyproline and polyglutamine stretches are associated with protein- and nucleic acid binding and transcription factor activity.[182,183] Protein segments enriched for glutamine and asparagine often occur in disordered regions[184] and are abundant in eukaryotic proteomes,[185] despite their propensity to aggregate or form coiled-coil structures.[186] The aggregation propensity of the Q/N-enriched segments has been exploited to mediate the formation of different, physiologically relevant assemblies such as P-bodies (e.g. Ccr4 and Pop2), stress granules and processing bodies,[187] but expanded polyglutamine repeats have also been associated with neurodegenerative disorders, the most well known being Huntington's disease.[188] Moreover, several prion-like yeast proteins (e.g. Sup35p and Ure2p) contain intrinsically disordered Q/N-rich protein segments that have been implicated in the switch between a soluble and an insoluble, aggregated, form.[185,189] Another example of functional disordered repeats occurs in the SR protein family of splicing factors (e.g. ASF/SF2 and SRp75).[190,191] SR proteins mediate assembly of spliceosome components and consist of an N-terminal RNA-recognition motif and a disordered C-terminus with tandem repeats of arginine and serine residues (RS domain). The RS domain is phosphorylated[192] and predicted to be largely disordered,[190] suggesting that the intrinsic disorder in the RS domain facilitates recruitment of spliceosome components. Other disordered repeats associated with specific function include repeats of lysine, alanine, and proline in the histone H1 C-terminal domain, which are involved in the formation of 30nm chromatin fiber by binding linker DNA between the nucleosomes.[193] Some of these repeats also have the propensity to undergo phase transition from a soluble monomeric state to an insoluble large assembly form (see **section 9.1.**)

6. Protein interactions

Disordered regions in the native unbound state exist as dynamic ensembles of rapidly interconverting conformations,[143,194] which can be described by an energy landscape.[195] Conditions, post-translational modifications and binding events change the energy levels of individual conformations as well as the relative energies between conformations.[196-198] As a result, the contribution of individual conformations to the protein ensemble changes under different conditions. For example, addition of specific binding partners can result in a population shift in the ensemble towards the conformation that is most favorable for binding; a phenomenon referred to as conformational selection.[99,123,196,198] This mechanism has been observed in both protein-protein and protein-nucleic acid interactions.[196]

Intrinsic disorder-mediated molecular interactions have been proposed to work using a combination of conformational selection and induced folding (i.e. when a protein undergoes a disorder-to-order transition upon association with its binding partner).[84,124] These mechanisms of binding are two extreme possibilities and are not mutually exclusive. Both play a role in the interaction between two proteins, the dominant mechanism depending for example on the concentrations of the individual proteins.[199] Evidence for the role of conformational selection in IDP binding comes from the interaction between PDEγ and the α-subunit of transducin,[200] which is important in phototransduction. The dynamic ensemble of unbound PDEγ includes a loosely folded state that resembles its structure when bound to transducin. Evidence for the role of induced folding, or coupled folding and binding, comes from a study investigating the disordered pKID region of CREB and the KIX domain of CREB-binding protein. Upon binding of pKID to the KIX domain, an ensemble of transient encounter complexes forms, which appear to be stabilized primarily by hydrophobic contacts and evolve to form the fully bound state via an intermediate state without disassociation of the two domains.[83,201]

6.1. Fuzzy complexes

Although disordered protein regions frequently fold upon interaction with other proteins, complexes with IDPs often retain significant conformational freedom and can only be described as structural ensembles.[202] The conformations that disordered proteins adopt in the bound state cover a continuum, similar to the structural spectrum of free, unbound IDPs,[203] and ranges from static to dynamic, and from full to segmental disorder.[202] In static disordered complexes, disordered regions can adopt multiple well-defined conformations in the complex, whereas in dynamic disorder they fluctuate between various states of an ensemble.

Disorder in the bound state can be classified into four molecular modes of action, each of which is associated with specific molecular functions (**Figure 11A-D**).[202,204] (i) The polymorphic model is a form of static disorder, with alternative bound conformations serving distinct functions by having different effects on the binding partner. Examples are the Tcf4 catenin binding domain[205] and the WH2 binding domain of e.g. thymosin β4 and Ciboulot,[206] which have been shown to adopt several distinct conformations upon β-catenin and actin binding, respectively. Different actin-WH2 domain complexes have alternative interaction interfaces and result in actin polymers with different topologies.[206] The (ii) clamp and (iii) flanking models represent forms of dynamic disorder in which complex formation either involves folding upon binding of two disordered segments that are connected by a linker that remains disordered, or the reverse situation, respectively. The cyclin-dependent kinase (Cdk) inhibitor p21 for example acts as a clamp. It contains a dynamic helical subdomain that serves as an adaptable linker that connect two binding domains and enables these to specifically bind distinct cyclin and Cdk complex combinations.[207] In both the clamp and flanking models, disordered regions near the interacting protein segments (often short peptide motifs) contribute to binding by influencing affinity and specificity.[202,208] This phenomenon relates to the importance of the sequence context in modulating disordered binding elements (see **section 3.**). Finally, (iv) the random model is an extreme version of dynamic disorder in protein complexes, which occurs when the IDR remains largely disordered even in the bound state. In this case, interaction is achieved via linear motifs that do not get fixed upon binding. An example is the self-assembly of elastin, where solid-state NMR has provided evidence for dynamic disorder within elastin fibers as they exhibit random-coil like chemical shift values.[209] Another case is the complex between the Cdk inhibitor Sic1 and the SCF ubiquitin ligase subunit Cdc4, which is formed in a phosphorylation-dependent manner.[210] At any given time, only one out of nine Sic1 phosphorylation sites interact with the core Cdc4 binding site, while the others contribute to the binding energy via a secondary binding site or via long-range electrostatic interactions (**Figure 12O**). Hence, binding interchanges dynamically within the Sic1-Cdc4 complex to provide ultra-fine tuning of the affinity.[210]

Bound disordered regions can impact the interaction affinity and specificity of the complex and tune interactions of folded regions[204] with proteins or DNA.[211] Four different mechanisms have been proposed for the formation of fuzzy complexes (**Figure 11E-H**). (i) Conformational selection, when the disordered region shifts the conformational equilibrium of the binding interface towards the bound form. The fuzzy N-terminal tail of the Max transcription factor for example reduces electrostatic repulsion in the basic helix-loop-helix (bHLH) domain and thereby facilitates formation of the DNA recognition helices, which increases binding affinity by 10-100 fold.[212] (ii) The disordered region(s) modulate flexibility of the binding interface. The serine- and arginine-rich region of the Ets-1 transcription factor exemplifies this mechanism and reduces DNA binding affinity by 100-1000 fold.[213] (iii) Competitive binding of the disordered region. Here the IDR acts as a competitive inhibitor of other regions in the same protein for binding to a partner. The acidic fuzzy C-terminal tail of high-mobility group protein B1 (HMGB1) negatively regulates interaction of the HMG DNA binding domains by occluding the basic DNA-binding surfaces.[214] (iv) The disordered region serves to tether a weak-affinity binding region to increase its local concentration. For example, a fuzzy N-terminal domain anchors the human positive cofactor 4 (PC4) to herpes simplex virion protein 16 (VP16)[215] and other trans-activation domains. All mechanisms of disordered complex formation affect binding to different degrees and can be further tuned by post-translational modifications.[204,211]

6.2. Binding plasticity

Structural analysis of a large number of intrinsic disorder-based protein complexes resulted in another categorization of IDRs based on their binding plasticity (**Figure 12**).[216] Examples of relatively static IDR-based complexes are (i) mono- and polyvalent complexes, which typically show binding of disordered segments to one or multiple spatially distant binding sites on the interaction partners, respectively, (ii) chameleons, such as p53, that have different structures when binding to different proteins, (iii) penetrators that bury significant parts of the protein inside their binding partners, and (iv) huggers, which function in protein oligomerization, for example by coupled folding and binding of disordered monomers. In addition to these relatively static complexes involving IDRs, one can identify coiled-coil-based complexes. Regions that make up coiled coils are typically highly disordered in monomeric state and gain helical structure upon coiled-coil formation, giving rise to several distinguishable types of complexes, such as intertwined strings, connectors, armatures, and tentacles.

7. Evolution

Disordered regions typically evolve faster than structured domains.[50-55,92] This behavior largely stems from a lack of constraints on maintaining packing interactions, which drives purifying selection in structured sequences.[217] However, disordered residues do display a wide range of evolutionary rates (**Box 3**). The following section discusses the evolutionary classifications of disordered protein regions. IDRs with similar function and properties tend to have similar evolutionary characteristics.

7.1. Sequence conservation

While the primary amino acid sequence of disordered regions evolves at different speeds, the property of disorder is usually conserved for functional sequences.[53,137] Sequence conservation of IDRs varies according to their specific functions and provides another means for their classification.[53,218,219] Three biologically distinct classes of IDRs with specific function were identified using a combination of disorder prediction and multiple sequence alignment of orthologous groups across 23 species in the yeast clade (**Figure 13**): (i) flexible disorder describes regions where disorder is conserved but that have quickly evolving amino acid sequences (i.e. there is a requirement to be disordered, regardless of the exact sequence), (ii) constrained disorder describes regions of conserved disorder with also highly conserved amino acid sequences, and (iii) non-conserved disorder, where not even the property of being disordered is conserved in close species. For flexible disorder, low sequence conservation is expected if the property of disorder itself, as opposed to disorder in combination with specific sequence, is the only requirement for function. Examples of functions that mainly require the biophysical flexibility of disordered regions are entropic springs, spacers, and flexible linkers between well-folded protein domains.[36,38,56,57] The linker in RPA70 is a clear example where the dynamic behavior is conserved even when the sequence conservation is low.[59] Flexible disorder is the most common of the three evolutionary classes (just over half of disordered residues in yeast) and appears to account for many of the characteristics traditionally associated with disordered regions, such as strong association with signaling and regulation processes,[35,49,89,220-222] rapid evolution of primary sequence,[50-55,92] the presence of short linear motifs (which are itself conserved, see below),[46,71] and tight regulation (see **section 8.**).[67,223] By contrast, constrained disorder (about a third of disordered residues in yeast) is associated with different properties and functions, such as chaperone activity and RNA-binding ribosomal proteins.[53] Many proteins that contain the evolutionary constrained type of disorder can adopt a fixed conformation, suggesting that these regions might undergo folding upon binding to their targets. This structural transition might impose a high degree of local structural constraints, which results in constraints on the primary protein sequence alongside requirements to be flexible.[53] Constrained disordered residues also occur more often in annotated protein sequence families (domains) than flexible disorder, but both types are strongly depleted in domains compared to structured regions. In human, both flexible and constrained disorder are enriched in proteins with functions related to differentiation and development.[224] Finally, non-conserved disorder accounts for around 17% of disordered residues in yeast and appears to be largely non-functional.

Short linear motifs (see **section 3.1.**)[47,105] constitute a special case. Even though SLiMs almost exclusively lie within disordered regions, their own amino acid sequence tends to be conserved.[47] These properties, together with the difficulty of aligning swiftly evolving disordered sequences, leads to the impression that motifs ‘move around’ when comparing their position in different sequences. In many ways, the disordered regions that contain SLiMs constitute flexible disorder as by the above classification, as their main role is to provide flexibility to enable access to the linear motif for proteins that will bind them as ligands[225] or deposit post-translational modifications.[46,47] Phosphorylation sites are closely related to short linear motifs that function in binding, but are often too short and weakly conserved to recognize via computational means.[226] More than 90% of sites phosphorylated by the yeast Cdk1 are in predicted disordered regions,[66] as consistent with previous studies highlighting the importance of IDRs as display sites for phosphorylation and other PTMs.[44,45] Comparison of the phosphorylation sites in orthologs of the Cdk1 substrates revealed that the precise position of most phosphorylation sites is not conserved. Instead, clusters of sites move around in the alignment of rapidly evolving disordered regions.[68,227] Another example of the role of flexible disorder in signaling and regulation is the yeast serine-arginine protein kinase Sky1, which regulates proteins involved in mRNA metabolism and cation homeostasis. The Sky1 C-terminal loop is intrinsically disordered and contains phosphosites that are important for regulating its kinase activity.[228] Conservation analysis has shown that the loop is conserved for disorder but not for sequence.[53]

The combination of sequence conservation of IDRs and conservation of their amino acid composition between human and seven other eukaryotes (chimp, dog, rat, mouse, fly, worm and yeast) also identifies function preferences.[219] IDRs with high residue conservation (HR) are enriched in proteins involved in transcription regulation and DNA binding. Low residue conservation in combination with high conservation of the amino acid type composition (LRHT) of the IDR (i.e. high similarity of overall amino acid composition between the human IDR and its orthologs) is often associated with ATPase and nuclease activities. Finally, IDRs which show neither conservation of sequence nor of composition (LRLT) are abundant in (metal) ion binding proteins.

7.2. Lineage and species specificity

Increasingly complex organisms have higher abundances of disorder in their proteomes.[35] An average of 2% of archaeal, 4% of bacterial and 33% of eukaryotic proteins have been predicted to contain regions of disorder over 30 residues in length,[35] although there is much variation within kingdoms.[229,230] In human, 31% of proteins are highly unstructured[67] and 44% contain stretches of disorder longer than 30 residues.[139,173] Human IDPs are spread relatively uniformly across the chromosomes, with percentages ranging from 38% (for genes encoding IDPs on chromosome 21) to 50% on chromosomes 12 and X.[139] A computational analysis of disorder in prokaryotes has corroborated the higher abundance of disorder in Bacteria as compared to Archaea.[231] Moreover, in agreement with the low abundance of disorder in prokaryotes, none of the 13 mitochondrial-encoded proteins are disordered.[139] Systematic analysis of IDP occurrence in 53 archaeal species showed that disorder content is highly species-dependent.[232] For example, *Thermoproteales* and *Halobacteria* proteomes have 14% and 34% disordered residues, respectively. Harsh environmental conditions seem to favor higher disorder contents, suggesting that some of the archaeal IDPs evolved to help accommodate hostile habitats.[233] Furthermore, disorder is more common in viruses than in prokaryotes.[234] The characteristics of IDRs seem well suited for especially small RNA viruses with extremely compact genomes.[235,236] For example, disordered regions could buffer the deleterious effects of mutations introduced by low-fidelity virus polymerases better than structured domains would.[234] The flexibility of IDRs to interact with many different proteins, such as proteins of the host immune system, is another benefit for compact viruses since it maximizes the amount of functionality they encode while minimizing the required genetic information.[237]

In addition to the variation in prevalence of disordered regions between species, different kingdoms of life seem to use conserved IDRs for different functions: eukaryotic and viral proteins use disorder mainly for mediating transient protein-protein interactions in signaling and regulation, while prokaryotes use disorder mainly for longer lasting interactions involved in complex formation.[137] Thus, knowledge on the lineage, species and origin of a disordered region could help in predicting its likely function.

7.3. History and mechanism of repeat evolution

Tandem repeats are enriched for intrinsic disorder (**see section 5.7.**) and IDRs are increasingly abundant in increasingly complex organisms (**see section 7.2.**). The genetic instability of repetitive genomic regions in combination with the structurally permissive nature of IDRs might have driven the increase in the amount of disorder during evolution. Disordered repeat regions have been shown to fall into three categories, based on their evolutionary history and acquired functional properties (**Figure 14**):[60] type I regions have not undergone function diversification after repeat expansion (e.g. the titin PEVK domain), type II comprises repeats that acquired diverse functions due to mutation or differential location within the sequence (e.g. the C-terminal domain of eukaryotic RNA polymerase II), and type III regions gained new functions as a consequence of their expansion per se (e.g. the prion protein octarepeat region).

8. Regulation

Altered availability of IDPs has been associated with diseases such as cancer and neurodegeneration.[220,223,238-241] Indeed, genes that are harmful when overexpressed (i.e. dosage-sensitive genes) often encode proteins with disordered segments.[242] Multiple mechanisms at different stages during gene expression (from transcript synthesis to protein degradation) carefully control the availability of IDPs in order to minimize harmful effects.[67] Their tight regulation ensures that IDPs are available in appropriate levels and for the right amount of time, thereby minimizing the likelihood of ectopic interactions. Disease-causing altered availability of IDPs may result in imbalances in signaling pathways by sequestering proteins through non-functional interactions involving disordered segments (i.e. molecular titration[223]). The next section discusses possible functional roles of proteins with IDRs based on their cellular regulatory properties such as transcript abundance, alternative splicing and degradation kinetics.

8.1. Expression patterns

Five different expression patterns were identified for transcripts encoding highly disordered proteins by investigating the mRNA levels from over 70 different human tissues and comparing the number of tissues in which IDP transcripts are expressed against the level of expression (**Figure 15**).[173] The expression classes are associated with specific functions. (i) The first subgroup (**Figure 15**, light blue markers) shows constitutive high expression in all tissues and consists exclusively of large ribosomal subunit proteins, which are almost entirely disordered. (ii) The second group (blue-green) represents transcripts that show high expression levels in the majority of tissues. These often function as protease inhibitors, splicing factors and complex assemblers. (iii) Moderately expressed transcripts (green) typically encode disordered proteins involved in DNA binding and transcription regulation. (iv) IDPs that are expressed in a tissue-specific manner (yellow) are enriched for cell organization regulators, transcription cofactors and factors that promote complex disassembly. Finally, (v) the remaining transcripts form a group (grey) not detected to be abundant in any of the tissues studied. This low and transient expression group contains more than half of the IDP transcripts analyzed and has a variety of functions (**Figure 15**).

8.2. Alternative splicing

Trends in transcriptional regulation (alternative promotor and polyadenylation site usage) and post-transcriptional regulation (alternative splicing by inclusion or exclusion of exons) can also be informative of the role that specific disordered protein regions play in the cell (**Figure 16**). Alternatively spliced exons are overall more likely to encode intrinsically disordered rather than structured protein segments.[139,243-245] This tendency is even more pronounced in alternative exons whose inclusion or exclusion is regulated in a tissue-specific manner.[246] IDRs that are encoded by these tissue-specific alternative exons frequently influence the choice of protein interaction partners and can be instrumental in protein regulation[246,247] by embedding (i) binding motifs and (ii) residues that can be post-translationally modified,[246] however (iii) sole alternation in the length of a disordered region[248] can also modulate the overall protein function (**Figure 16**). Changes in IDR length can be an effective mechanism for modifying the affinity of interactions that a protein makes, particularly in instances where a disordered region is responsible for the positioning of protein binding motifs or domains.[249,250] Among the alternative exons, those that exhibit conserved splicing patterns across different species are particularly likely to have important regulatory roles. For example, tissue-specific exons, which are alternatively spliced in multiple different mammals, remarkably often contain IDRs with embedded phosphosites.[251] Disordered regions these exons encode are hence likely to act as modulators of protein function depending on the tissue of gene expression.[251] While tissue-specific exons that are alternatively spliced in a conserved fashion often code for phosphosites, the emergence of novel exons in a gene, although at first likely detrimental[252] is a possible template for the evolution of short interaction motifs.[253] Furthermore, changes in exon regulation can also be important for the emergence of novel adaptive functions. Accordingly, protein segments encoded by exons, which are alternatively spliced either in a single species or in a whole evolutionary lineage, are enriched in short binding motifs and alternative inclusion of disordered regions encoded by these exons is conceivably a source of evolutionary novelty.[254]

In addition to the tendency of cassette alternative exons to frequently encode IDRs, exons adjacent to the alternatively spliced ones are also likely to code for disordered regions around the insertion point for the alternative segment.[224,244] These disordered regions provide the structural flexibility that tolerates both presence and absence of the alternatively spliced segment, but they can also contain interaction motifs themselves.[224] Furthermore, on the transcriptional level, diversity in protein isoforms can be created through both alternative splicing and usage of alternative promoters and polyadenylation sites. Protein segments that are encoded by the two latter mechanisms can contain disordered regions with motifs that define protein localization and stability.[255] Taken together, these examples illustrate how better understanding of gene regulation and knowledge of evolutionarily conserved and novel patterns in transcript processing can provide insights into possible functional roles of whole proteins, but also individual protein regions.

8.3. Degradation kinetics

Another emerging functionality of disordered regions is their role in protein degradation.[256-263] For example, protein half-life weakly correlates with the fraction of disordered residues[67,259] and proteins that get ubiquitinated specifically upon heat shock stress are generally disordered.[264] Although ubiquitination by E3 ligases has a dominant role in recruiting proteins to the proteasome for degradation,[265,266] some IDRs of sufficient length seem to allow for efficient initiation of degradation by the proteasome independent of the ubiquitination status, as supported by *in vitro* experiments showing that degradation of tightly folded proteins is accelerated when a disordered region is attached to the substrate.[257,263] Efficient degradation only occurs when the disordered terminal region is of a certain minimal length,[263] but degradation may be initiated by IDRs either at the protein terminus or internally.[256-263] Proteins that contain IDRs of sufficient length may therefore have increased turnover, although the exact length requirements will depend on the substrate. At the same time, not all IDRs influence protein half-life. For example disordered polypeptides with specific amino acid compositions can attenuate rather than accelerate degradation by the proteasome.[267-269] The formation of proteins complexes or transient interactions with other proteins may also protect IDPs from degradation. Thus, we can distinguish a novel function class of IDRs: those that influence protein degradation (degradation accelerators) versus those that do not. These properties might be associated with specific protein function. For example, proteins that contain IDRs of a given length are probably more susceptible to degradation, possibly linking them to functions of IDPs with low expression, although this is not necessarily the case.

Some highly disordered proteins (e.g. p53, p73, IκBα, BimEL) can, at least *in vitro*, be degraded by the 20S proteasome independent of ubiquitination.[270-275] Specialized proteins termed ‘nannies’ have been shown to bind to and protect IDPs from ubiquitin independent 20S proteasomal degradation.[276] A free IDP, such as newly synthesized p53, might be degraded by the 20S proteasome, which leads to fast degradation kinetics. After a nanny binds the IDP (Hdmx in the case of p53), slower, ubiquitin-dependent degradation by the 26S proteasome takes place. This biphasic decay has been proposed as a way to distinguish structured proteins from IDPs and the proteins that protect them from degradation.[276]

9. Biophysical properties

9.1. Phase transition

The involvement of IDRs in phase transitions is a surprising finding that provides a biophysical angle to the characterization of proteins that harbor disordered regions. Li and co-workers[116] observed that interactions between recombinant proteins that contain multiple copies of an SH3 domain and IDRs with multiple instances of the proline-rich SH3 interaction motif (see **section 3.1.**) produced sharp liquid–liquid-demixing phase separations that result in micrometer-sized liquid protein-based droplets (**Figure 17A**). The concentrations needed for the phase transition depend on the valency of the interacting elements. Importantly, experiments with the natural NCK–nephrin–N-WASP (neuronal Wiskott–Aldrich syndrome protein) complex, which contains multiple copies of the same SH3 interaction partners, showed the formation of similar dynamic droplets that lead to a significant increase in the activity of the actin nucleation factor Arp2/3.[116] The formation of the droplets is controlled by the degree of phosphorylation of one of the interaction partners, which potentially explains how the phase transition is regulated in the cell.

A related phenomenon occurs with RNA-binding proteins that contain IDRs with low sequence complexity, which have been associated with the regulated formation of cellular RNA granules.[277] Various types of cellular RNA granules are used to modulate the fate of specific mRNAs, but their assembly mechanism has remained a mystery. Kato and co-workers[278] reconstituted granule-like RNA assemblies *in vitro* by exploiting low complexity IDRs. They demonstrated that the low-complexity IDRs of certain RNA-binding proteins were necessary for the formation of granule-like assemblies and that high concentrations of these regions lead to a reversible phase transition with a highly dynamic hydrogel state (**Figure 17B**). Interestingly, hydrogels formed by the low-complexity IDR of one purified member of the granules are capable of binding IDRs of other members and thereby enable the assembly of heterogenous macromolecular structures.[278] Overall, these findings indicate that the biophysical properties of certain IDRs (such as those that contain specific low-complexity regions or linear motifs) enable phase transitions that are likely to be exploited in various macromolecular assemblies and could function to bridge the length scale of proteins with that of organelles.[279]

10. Discussion

In order to get closer to full understanding of living cells, we need to know the function of each of their elements. The human genome project and the many sequencing projects since have helped reveal the number and make-up of the genes. Experimental research focused on understanding how individual proteins work on the molecular level has enabled enormous progress in our understanding of the workings of proteins in general and of the systems they work in. However, the majority of studies investigate a minority of individual proteins, which are interesting for a variety of reasons, such as their relevance for disease or because they are classical study objects. Thus, still many genes and the proteins they encode have not been studied and have unknown function.

Many of the functionally uncharacterized proteins will be similar to already characterized ones.[8-10] This notion forms the basis for computational methods that aim to improve annotation coverage by predicting the function of novel and undefined proteins based on information from better-studied proteins. Databases such as Pfam[22] and SCOP[24] attest to the success of these approaches. However, existing methods are focused primarily on sequences that give rise to well-folded protein structures and domains. As a result, it is much harder to gain insight into function of intrinsically disordered proteins (IDPs) and regions (IDRs), despite the increasing evidence of their prevalence and importance for protein functionality (**Figure 1**).[49] Many important disease proteins such as p53, Myc, α-synuclein and BRCA1 are highly disordered, underscoring the importance of disordered regions for understanding the molecular basis of human diseases.[223,238]

In this review, we have assembled an overview of the major approaches used to classify and categorize IDPs and IDRs (**Table 1**). These classification schemes help us understand how disordered protein functionality is defined and could be used to enhance function prediction for disordered protein regions. In these final sections we discuss the resources that are currently available to gain insight into IDR function, we address potential areas of improvement of the current approaches, and we speculate how combinations of multiple existing classifications could achieve high quality function prediction for IDRs. Finally, we suggest areas where increased efforts are likely to advance our understanding of the functions of structural disorder in proteins.

10.1. Current methods for function prediction of IDRs and IDPs

What methods and resources can a researcher use to gain insight into the functions of the disordered regions in a protein? Current approaches are mainly based on the presence of functional features such as short linear motifs (SLiMs), molecular recognition features (MoRFs), and intrinsically disordered domains (IDDs) (see **section 3.**). These aspects have the potential to shed light on which interaction partners an IDR may have and how many, as well as the mode of binding.

10.1.1. Linear motif-based approaches

Mapping of well-characterized linear motifs onto other protein sequences holds particular promise for discovering novel functionality. For example, proteomic characterization of the motif (RxxPDG) that recruits Tankyrase ADP-ribose polymerases has lead to the identification of novel Tankyrase substrates and explains the basis for mutations causing cherubism disease.[280] Similarly, proteome-wide searches for the SxIP motif have resulted in the identification of previously uncharacterized microtubule plus-end tracking proteins.[281] However, these types of individual studies require considerable resources.

ScanSite,[282] MiniMotif,[106] and ELM[105] are three major efforts aimed at the annotation of known instances of linear motifs, which are primarily found in IDRs, and their binding partners. ScanSite primarily identifies linear motifs that are likely to be phosphorylated and play key roles in signaling, such as SH2 and 14-3-3 motifs.[282] Annotation of these sequence motifs is based on results from binding experiments with peptide libraries and phage display experiments. The MiniMotif[106] and ELM[105] databases aim to categorize linear motifs of all functions based on in-depth manual annotation of experimentally validated instances from the literature. Although these resources are excellent repositories of the functional sites that occur in IDRs, they do have certain shortcomings. For example, ScanSite has a limited scope in identifying phosphorylation sites and the annotations from MiniMotif are not publicly available. Though the ELM database is the most comprehensive database of functional features within IDRs, at present it does not have the resources to annotate all motifs in the literature; ELM contains almost 200 classes of linear motifs with over 2000 instances but more than 250 classes await annotation with this number constantly increasing. This has meant ELM is limited to annotating (a fraction) of the shorter motif classes and ignores the longer binding modules in disordered regions.

Complementary to the annotation efforts, the linear motif resources employ prediction methods that map functionality onto regions of proteins with unknown function (i.e. unannotated regions). For example MiniMotif and ELM use regular expressions derived from experimentally validated and curated motif instances to search protein sequences. These searches bring up functional descriptions of sequence instances that match the regular expressions. A major problem in the computational detection of short motifs in particular is the high false positive rate, which means that it is very difficult for users to identify the instances that are most likely to be functional from the large total of mostly non-functional motif instances that result from these searches. To overcome this issue, both databases have developed additional methods to improve prediction accuracy that rely on the use of additional context information, such as accessibility (using structural models[283] and predictions of intrinsic disorder[71]), evolutionary conservation,[284,285] cell compartment (based on annotation),[106,286] and protein-protein interactions.[108,287,288] These efforts will need to be combined in the future with a clearer user interface so researchers can more easily identify the most relevant instances.

De novo predictors make up the final category of motif resources. These predictors computationally identify putative uncharacterized motifs in protein sequences. There are two broad types: predictors that identify clusters of amino acids that are more conserved than surrounding residues (e.g. SLiMPrints[289]) or those that find short peptide patterns that are over-represented in a set of sequences (e.g. DiliMot[290] and SLiMFinder[291]). Although both approaches have been combined with the gene ontology terms of the identified proteins, further development is required to define potential functionality.

10.1.2. Molecular recognition feature-based approaches

Two important methods exist for identifying novel binding modules in IDRs based on concept of molecular recognition features (MoRFs). MoRFpred predicts sequences that undergo disorder-to-order transitions of all types of MoRFs (α, β, coil and complex) using a combination of sequence alignment and machine learning predictions based on amino acid properties, predicted disorder, B-factors, and solvent accessibility.[292] ANCHOR also predicts parts of disordered regions that are likely to fold upon binding with their interactors, but does so by identifying segments that cannot form enough favorable intrachain interactions to fold on their own and are likely to gain stabilizing energy by interacting with a globular protein partner.[293,294]

An important shortcoming of the MoRF predictions is the difficultly in identifying which of the binding sites are relevant and what their functionality might be. This is primarily because the results are not linked to known MoRF instances with annotated functions, as is the case for linear motifs, and no clues are provided regarding the potential role of a binding site or its interacting partners. The IDEAL database[295] collects verified elements in disordered regions that undergo coupled folding and binding upon interaction (**Box 1**). The careful annotation of well-described MoRFs in terms of their primary sequence propensities or interaction interfaces as well as their known binding partners, and integration of these annotations with MoRF predictions, would likely improve the use of these predictions for gaining insight into IDR functionality.

10.1.3. Intrinsically disordered domain-based approaches

Few attempts have systematically annotated protein domains that are largely made up of intrinsic disorder. To a certain extent, Pfam[22] has developed models to predict intrinsically disordered domains (e.g. KID, WH2, RPEL and BH3 domains). However, this seems to be a simple consequence of the fact that these disordered domains can be described and detected by sequence profiles, rather than an effort directed at annotating long IDRs. ELM[105] has also annotated a small number of long disordered domains, such as the WH2 motif, however the main focus of the database remains on short motifs. Finally, some of the IDRs that are present in annotated domains are in fact MoRFs or linear motifs, and linear motifs also frequently fold upon binding like MoRFs, underscoring the underlying connections between linear motifs, MoRFs, and IDDs as functional elements (see **section 3.4.**).

10.1.4. Other approaches

Only a few IDR classifications that are not based on linear motifs, MoRFs or IDDs have so far been exploited for function prediction. FFPred is a correlation-based approach that uses the length and position of IDRs along a sequence (see **section 5.5.** and **5.6.**), among other general protein features, to predict the function of the protein in terms of gene ontology categories (molecular activities and biological processes).[176,296-298] The DisProt database of protein disorder[166] (**Box 1**) lists functions of individual disordered regions, where know from experiments. The major limitation here being the small number of regions for which exact function has been characterized.

10.2. Requirement for annotation

Future effort in the classifications of IDRs and IDPs must be directed at annotation. Substantiating classes with more examples will lead to refinement of their function descriptions. For example, there are only a limited number of well-characterized examples of proteins that contain the evolutionary flexible (e.g. RPA70 and Sky1) or constrained types of disorder (Rpl5 and Hsp90). The same is true for the different classes of dynamic disorder in protein complexes, though efforts are ongoing there.[204] In terms of the functional features of IDRs, there is a need for annotating MoRFs and longer disordered binding regions as described in the previous section. Efforts directed at short linear motif have been very successful, but only a small fraction of the potentially thousands of motifs[299] have been annotated. Pfam contains almost 15,000 curated protein families,[22] while ELM contains less than 200 motif classes,[105] suggesting that significant numbers of functional features are still to be identified and further annotation is required. High-quality resources that collect all the experimentally validated functional regions of intrinsically disordered regions will provide a strong basis to map functional features onto novel proteins of unknown function.

10.3. Integration of methods for finding IDR and IDP function

The current methods for finding and classifying IDR and IDP function have been successful in the area of their focus. However, not all functional characteristics of disordered regions have been fully exploited, and neither is there a resource that brings all of these aspects together. The combination of multiple categorizations and features of IDRs is likely to provide a better understanding of the functionalities encoded in these regions.

A comprehensive IDR function resource should have several aspects. It starts with a reliable consensus disorder prediction for the protein sequence of interest, such as available in the D^2P^2 database (**Box 1**).[48] Functional features, such as SLiMs (see **section 3.1.**), MoRFs (see **section 3.2.**), and disordered domains (see **section 3.3.**) can then be mapped on every disordered part of the protein. The disorder profile allows for the identification of individual IDRs in the protein, as well as the calculation of disorder properties of the whole protein, such as which disorder predictors support which IDRs (see **section 5.2.**), the overall degree of disorder (see **section 5.4.**), the length of the individual disordered regions (see **section 5.5.**), or the amount of disorder at the termini (see **section 5.6.**). These can be used to assign general function to the proteins, such as gene ontology terms that correlate with these properties. Patterns in primary amino acid sequence could reveal additional function. For example, the presence of tandem repeats (see **section 5.7.** and **7.3.**) may point at involvement in certain processes. The overall sequence composition and the distribution of charges (see **section 5.1.**) could indicate conformational properties such as the degree of compaction of the chain (see **section 4.**) and the combination of primary sequence complexity and disorder propensity could suggest function as well (see **section 5.3.**).

Integration of other types of information will determine what classifications can additionally be used. Addition of domain information, such as Pfam, can provide insight into the role of disordered segments that are commonly associated with specific structured domains (see **section 3.3.**). Protein-protein interactions and structures of protein complexes could indicate interacting partners of IDR binding elements and the mode of interaction (see **section 6.**). Information about sequence conservation (see **section 7.1.**) is another important aspect and could provide clues about evolutionary constrained or flexible types of disorder, which are implicated in different types of functioning. Knowledge on the origin of a disordered region in evolution or the species containing the protein sequence of interest suggests possible roles as well (see **section 7.2.**). Furthermore, data describing regulatory properties like gene expression levels (see **section 8.1.**), alternative splicing (see **section 8.2.**) and degradation kinetics (see **section 8.3.**) could implicate IDRs in regulating protein availability and may suggest or reject roles as for example interactions hubs. Finally, biophysical properties of the protein, such as the potential of multivalent elements to undergo phase transitions (see **section 9.1.**), may suggest involvement in spatial-temporal localization of cellular components for example through the construction of subcellular granules.

The hypothetical resource might be able to suggest function for some of the following examples, although it is clear that in other cases the biology will be too complicated and the outlook of function prediction as described here will be unrealistic. Therefore, these examples should at this point be considered as speculative. A long (more than 30 residues) IDR that shows signs of evolutionary flexible disorder and contains no short motifs or other predicted binding regions could be a flexible linker between domains or an entropic chain. A region containing a PxxPx[KR] motif flanked by evolutionary flexible disorder that is likely to retain an open conformation in the unbound form (based on the primary sequence) probably binds a class II SH3 domain, and might be involved in transcription processes if the IDR constitutes the C-terminus of a protein with an otherwise small degree of disorder. Long IDRs that are encoded by alternatively spliced exons and have several non-overlapping functional motifs and MoRFs might be part of signaling hubs or assemble multiprotein complexes, the type of which might be inferred from the combination of binding sites present. A constitutively expressed, largely disordered IDP with an amino acid composition promoting intrinsic coil conformations and conservation of both primary and disorder sequence is likely to be a ribosomal protein.

It is clear that some classifications will provide more useful and direct information about function than others. Some classifications have been proposed to contrast IDPs with structured proteins, which does not necessarily make them useful for a detailed description of disorder function per se. Others have limited use for prediction because they are conceptual only, or because of overlap in the properties they describe with other schemes. Moreover, not all approaches can realistically be incorporated in a tool. Binding functionality and sequence-based predictions will generally be possible, but predictions based on other types of data may be harder. For example, assignment of evolutionary constrained or flexible disorder requires automatic alignment of primary and disorder sequences, while gene expression subtypes can be derived from the wealth of microarray and RNA sequencing data. Various types of information are already brought together in the D^2P^2 database,[48] which contains information on disordered regions, MoRFs, PTM sites, and structured domains, and in ELM,[105] which shows information on linear motifs, disorder, phosphorylation, domains, protein-protein interactions and secondary structure. Further extension of resources like these, with information on both structured and disordered regions, holds great promise towards creating comprehensive overviews of the functional elements and properties of a protein.

10.4. Future directions

A major area of improvement in the description of disordered protein regions pertains to their dynamic behavior. IDRs fluctuate rapidly over an ensemble of heterogeneous conformations (see **section 4.1.**), the individual energy levels and propensities of which are determined by the primary sequence. The relationship between sequence and ensemble is important because it describes what part of the time the chain is in a compact state, and what part of the time it is more accessible (see **section 4.** and **5.1.**). Knowledge about these structural subtypes and about how sequence contexts and chemical modifications of the chain (e.g. by PTMs) modulate the ensemble is vital for correct description of IDR behavior and has direct implications for the functional roles such regions can have in the cell.

Existing methods are not designed to take structural dynamics into account. For example, current disorder prediction technology is successful at distinguishing sequence stretches that are likely to be disordered versus those that are likely to be part of autonomously folded domains, resulting in a binary verdict (disordered versus structured) within a certain confidence limit. Detailed prediction of conformational subtypes requires a more sophisticated description of disorder. For example, high-throughput atomistic simulations of sequence ensembles can provide information about the degree of conformational heterogeneity,[300] which can be quantified by various parameters, such as an information theory measure[301] or an order-parameter-like measure.[302] One could imagine a multiple-component scheme describing structural and dynamic characteristics that would assign for example residues in a random coil small values for the fractional population of secondary structure, a large value for spatial fluctuations, a fast interconversion rate, and large values for structural heterogeneity. Conversely, molten globule residues would be assigned a relatively large value for the fractional population of secondary structure, a smaller value for spatial fluctuations and structural heterogeneity, and a slower interconversion rate. There is considerable room for growth at the interface between atomistic simulations, physical theories, machine learning methods, and experiments, to enable the unmasking of the connection between disorder dynamics and molecular and system level functions of IDRs and IDPs.

Full understanding of the cellular function of IDPs will also require knowledge of their abundance, their interactions, and their physical state in the physiological context. Are IDPs always bound to target proteins, are they chaperoned, or are there pools of unbound IDPs? Answers to these questions will vary amongst different IDPs, but the discovery of features that can help classify and categorize them in terms of their cellular status will lead to more insights into their function. For example, entropic chains may mostly be disordered even in the cell, whereas effectors and assemblers may mostly be associated with other proteins in folded conformations and exchange binding partners by competition rather than by dissociation to the free, disordered state. Scavengers likely populate both disordered and ordered states, depending on whether or not their ligand is bound. Thus, investigations of the in-cell status of IDPs will be crucial towards understanding their biological roles.

11. Conclusion

The functional versatility of intrinsically disordered regions in proteins is remarkable. Our hope is that the overview of different groups, categories, types, and classes of IDRs in this review provides a basis for understanding how this functional versatility is achieved and that it offers novel ways of combining this knowledge to gain insight into the function of uncharacterized proteins.

Finally, we would like to stress that it is not all about intrinsic disorder. This review has focused on classifications for intrinsically disordered regions and proteins, because function annotation for these regions is lagging behind annotation of structured regions. However, proteins are modular[303] and their functional regions can be structured or disordered, or somewhere in between. The synergy between these fundamental building blocks of proteins leads to combinatorial diversity of functions and understanding how they work together will be key to understanding the full extent of protein function.

Acknowledgements

We thank our colleagues for stimulating discussions and Guilhem Chalancon for critically reading the manuscript and providing helpful comments. Funding: MRC (MC_U105185859); HFSP (RGY0073/2010) and the Virgo consortium, funded by the Dutch government project number FES0908, and by the Netherlands Genomics Initiative (NGI) project number 050-060-452 (RvdL); the American Cancer Society (RSG-07-289-01-GMC) and the National Science Foundation

(MCB-0939014) (GWD); the Momentum program (LP2012-41) of the Hungarian Academy of Sciences and the Hungarian Science Fund (OTKA NN 106562) (MF); the Odysseus grant G.0029.12 from the Research Foundation Flanders (FWO) (PT); the National Cancer Institute of the National Institutes of Health (CA096865) and the Skaggs Institute for Chemical Biology (PEW); the EMBO YI Programme and HFSP (RGY0073/2010) (MMB).

Boxes

Box 1: Databases of intrinsically disordered regions and proteins

Several resources exist that collect experimental or computational information on disordered regions in proteins. The Database of Protein Disorder (DisProt, http://www.disprot.org/) was developed to facilitate research on protein disorder by collecting and organizing the rapidly increasing knowledge about the experimental characterization and the functionalities of IDRs and IDPs.[166,304] For each disordered protein (i.e. a protein that contains at least one experimentally determined disordered region), the database includes the name of the protein, various aliases, accession codes, amino acid sequence, location of the disordered region(s), and methods used for disorder characterization. Additionally, where known, entries list the biological function of each disordered region and how each region of disorder is used for function. The data on disordered regions in DisProt often serves as a standard for the training and verification of disorder predictors. As of release 6.02 (May 24, 2013) DisProt contains 694 intrinsically disordered protein entries and 1539 disordered regions.

The IDEAL database (Intrinsically Disordered proteins with Extensive Annotations and Literature, http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/) also collects annotations of experimentally verified intrinsically disordered proteins.[295] This database focuses on the functional regions within IDRs, such as those that interact with other proteins and post-translational modification sites. IDEAL collects verified IDRs that undergo coupled folding and binding upon interaction (regions for which there is evidence for both a disordered isolated state and an ordered bound state), such as MoRFs and certain linear motifs (see **section 3.**). It also suggests putative sequences for which there is only evidence of an ordered bound state, but that are thought to undergo induced folding based on circumstantial evidence, such as the presence of a verified folding-upon-binding element in a homolog. The version of 30 August 2013 contains 340 proteins with annotated disordered regions of which 148 contain verified or putative elements that undergo folding upon binding.

Another database that collects experimental data and predictions of disorder in proteins is MobiDB (http://mobidb.bio.unipd.it/).[305] MobiDB collects experimental data from DisProt,[166] IDEAL,[295] and the Protein Data Bank[125] (missing residues in crystal structures and structurally mobile regions in NMR ensembles). Prediction data comes from three different methods. The total of disorder information is summarized in a weighted consensus. Version 1.2.1 (August 28, 2012) contains 26,933 proteins for which there is experimental data on the presence or absence of disorder and 4,662,776 proteins from 297 proteomes for which there are disorder predictions.

Finally, the Database of Disordered Protein Prediction (D^2P^2, http://d2p2.pro) stores disorder predictions made by nine different predictors for a large library of proteins from completely sequenced genomes.[48] Alongside the disorder predictions, it contains information on MoRFs (ANCHOR[293]), PTM sites (PhosphoSitePlus[306]), and domains (SCOP[24] and Pfam[22]). As of September 2013, D^2P^2 contains disorder predictions for 10,429,761 sequences in 1,765 genomes from 1,256 distinct species.

Box 2: Experimental detection of disordered regions

IDRs have been studied using a variety of experimental techniques, including NMR, X-ray, SAXS, proteomics, and protease susceptibility essays. NMR spectroscopy is the key method to characterize protein disorder, owing to its ability to provide residue-level information on protein structure and dynamics in solution.[307] Many aspects of disorder biology can be detected directly using NMR, including local disorder, folding upon binding, and disorder in complex. In contrast to NMR methods, detection of disorder using X-ray crystallography techniques is mainly indirect as it relies on missing electron density.[32] Another powerful method for detecting and characterizing IDPs is small-angle X-ray scattering (SAXS), which assesses protein dimensions and shape by measuring the scattered X-ray intensity caused by a sample. SAXS can be used to determine hydrodynamic parameters and the degree of globularity of a protein, which are good indicators to determine whether a protein is compact or unfolded.[146] High-throughput proteomic approaches are also used for identification of IDPs. These techniques enrich cellular extracts for disordered proteins, and then separate structured from disordered proteins, followed by identification (e.g. by mass spectrometry). For example, heat treatment enriches cell extracts for IDPs and depletes for proteins containing folded domains.[174] IDPs can also be identified on the basis of their susceptibility to degradation by the 20S proteasome under conditions in which structured proteins are resistant (see **section 8.3.**).[274] The degradation assays can also be used to identify binding partners of IDPs that provide protection against degradation. The DisProt, IDEAL and MobiDB databases collect experimentally verified disordered regions and proteins (**Box 1**).

Box 3: Evolution of disordered regions

IDRs generally evolve faster than their structured counterparts.[50-55,92] Direct comparison between the rates of evolution of structured and disordered regions in 26 protein families has shown that not all disordered regions evolve faster than structured regions; some evolve more slowly, and in other cases they evolve at roughly the same rate.[50] In order to get more insight into the evolution of disordered regions and proteins, we predicted intrinsic disorder in the human proteome using MULTICOM-REFINE.[308] We integrated the disorder status of the protein residues with their evolutionary rates across multiple sequence alignments of homologous proteins from 53 mostly vertebrate species in Ensembl Compara[1], calculated using the Rate4Site program.[309] As observed previously,[310] protein residues that are predicted to be disordered generally evolve more quickly (i.e. have much higher evolutionary rates) than those in structured regions (**Figure Box 3**, P value < 10^{-15}, Mann-Whitney U test). However, the distributions of evolutionary rates for disordered and structured residues are wide and overlap (**Figure Box 3**). Thus, although IDRs generally evolve more quickly, some disordered residues are conserved. In line with this, it has been shown that particular residue types, such as leucine, tyrosine, tryptophan, and proline, are more conserved in IDRs than other residue types.[52] Conserved residues and elements in IDRs are potentially important for function and might for example be part of protein-protein interaction interfaces and peptide motifs (see **section 7.1.**).

Tables

Table 1: Classifications of Intrinsically Disordered Regions and Proteins.

Basis for classification		Classes	Description	Examples
Function[33,38,56,57]		• Entropic chains	IDRs carrying out functions that benefit directly from their conformational disorder, e.g. flexible linkers and spacers.	MAP2 projection domain, titin PEVK domain, RPA70
		• Display sites	Flexibility of IDRs facilitates exposure of motifs and easy access for proteins that introduce and read PTMs.	p53, histone tails, p27, CREB kinase-inducible domain
		• Chaperones	Their binding properties (many different partners, quick interactions, and folding upon binding) make IDPs suitable for chaperone functions.	hnRNP A1, GroEL, Hsp33
		• Effectors	Folding upon binding mechanics allow effectors to modify the activity of their partner proteins.	p21, p27, calpastatin, WASP GTPase-binding domain
		• Assemblers	Assembling IDRs have large binding interfaces that scaffold multiple binding partners and promote the formation of higher-order protein complexes.	ribosomal proteins L7, L12, Tcf 3/4, CREB transactivator domain, Axin
		• Scavengers	Disordered scavengers store and neutralize small ligands.	chromogranin A, casein, pro-rich glycoproteins
Functional features	Linear motifs[46,105]	• Structural modification	Sites of conformational alteration of a peptide backbone.	peptidylprolyl cis-trans isomerase Pin1 sites
		• Proteolytic cleavage	Sites of post-translational processing events or proteolytic cleavage scission sites.	Caspase-3/-7, Seperase, Taspase1 scission sites
		• PTM removal/ addition	Specific binding sequences that recruit enzymes catalyzing PTM moiety addition or removal.	cyclin-dependent kinase phosphorylation site, SUMOylation site, N-glycosylation site
		• Complex promoting	Motifs that mediate protein-protein interactions important for complex formation often associated with signal transduction.	proline-rich SH3-binding motif, cyclin box, pY SH2-binding motif, PDZ-binding motif
		• Docking	Motifs that increase the specificity and efficiency of modification events by providing an additional binding surface.	KEN box degron, MAPK docking sites
		• Targeting or trafficking	Signal sites that localize proteins within particular subcellular organelles or act to traffic proteins.	nuclear localization signal, clathrin box motif, endocytosis adaptor trafficking motifs
	Molecular recognition features (MoRFs)[101]	• Alpha	Disordered motifs that form α-helices upon target binding.	p53 ~ Mdm2, p53 ~ RPA70, p53 ~ S100B(ββ), RNase E ~ enolase, inhibitor IA3 ~ proteinase A
		• Beta	Disordered motifs that form β-strands upon target binding.	Rnase E ~ polynucleotide

				phosphorylase, Grim ~ DIAP1, pVIc ~ Adenovirus 2 proteinase
		• Iota	Disordered motifs that form irregular secondary structure upon target binding.	p53 ~ Cdk2-cyclin A, Amphiphysin ~ α-adaptin C
		• Complex	Disordered motifs that contain combinations of different types of secondary structure upon target binding.	Amyloid β A4 ~ X11, WASP ~ Cdc42
	Intrinsically disordered domains (IDDs) [136,137]		Some proteins domains identified using sequence-based approaches are fully or largely disordered.	WH2, RPEL, BH3, KID domains
	Co-occurrence of protein domains with disordered regions [139,140]		Particular disordered regions frequently co-occur in the same sequence with specific protein domains.	
Structure	Protein quartet [32,34,144]	• Intrinsic coil	Flexible regions of extended conformation with hardly any secondary structure. High net charge differentiates these from disordered globules.	ribosomal proteins L22, L27, 30S, S19, prothymosin α
		• Pre-molten globule	Disordered protein regions with residual secondary structure, often poised for folding upon binding events. Lower net charge makes them more compact than coils.	Max, ribosomal proteins S12, S18, L23, L32, calsequestrin
		• Molten globule	Globally collapsed conformation with regions of fluctuating secondary structure.	nuclear coactivator binding domain of CREB binding protein
		• Folded	Structured proteins with a defined three-dimensional structure.	most enzymes, transmembrane domains, hemoglobin, actin, etc.
Sequence	Sequence-ensemble relationships [144,167]	• Polar tracts	Sequence stretches enriched in polar amino acids often form globules that are generally devoid of significant secondary structure preferences.	Asn-, and Gly-rich sequences, Gln-rich linkers in transcription factors and RNA-binding proteins
		• Polyelectrolytes	Amino acid compositions biased toward charged residues of one type; strong polyelectrolytes (high net charge) form expanded coils.	Arg-rich protamines, Glu/Asp-rich prothymosin α
		• Polyampholytes	Sequences with roughly equal numbers of positive and negative charges. Conformations of polyampholytes are governed by the linear distribution of oppositely charged residues, with segregation of opposite charges leading to globules, while well-mixed charged sequences adopt random-coil or globular conformations, depending on the total charge.	RNA chaperones, splicing factors, titin PEVK domain, yeast prion Sup35
	Prediction flavors [170]	• V	Predicted best by the VL-2V predictor, for which the hydrophobic amino acids are the most influential attributes.	*E.coli* ribosomal proteins
		• C	VL-2C is the best predictor for flavor C, which has more histidine, methionine and alanine residues than the other flavors.	poly- and oligosaccharide binding domains
		• S	Flavor with less histidine than the others, best predicted by predictor VL-2S, which has a measure of sequence complexity as the most important attribute.	proteins that facilitate binding and interaction

	Disorder-complexity [171]		IDPs from different function classes show distinct disorder-complexity distributions.	Disordered linkers of structured domains populate compact and disordered DC regions.
	Overall degree of disorder [35,50,67,139,173,174]	• Fraction	Categorization of proteins based on the fraction of residues predicted to be disordered.	0-10/10-30/30-100% disorder
		• Overall score	Overall disorder scores for the whole protein.	residue score average >0.5
		• Continuous stretches	Presence or absence of continuous stretches of disordered residues.	typically >30 residues
	Length of disordered regions [176]	• >500 residues	Proteins that contain disordered regions of different lengths are enriched for different types of functions.	transcription
		• 300-500 residues		kinase and phosphatase functions
		• <50 residues		(metal) ion binding, ion channels, GTPase regulatory activity
	Position of disordered regions [176]	• N-terminal	Proteins that contain disordered regions at different locations in the sequence are enriched for different types of functions.	DNA-binding, ion channel
		• Internal		transcription regulator, DNA-binding
		• C-terminal		transcription repressor/activator, ion channel
	Tandem repeats [180,181]	• Q/N	Glutamine- and asparagine-rich proteins regions are both important for normal cellular function and prone to harmful aggregation.	huntingtin, Sup35p, Ure2p, Ccr4, Pop2
		• S/R	Tandem repeats composed of arginine and serine residues are phosphorylated and disordered, and play a role in spliceosome assembly.	ASF/SF2, SRp75
		• K/A/P	Tandem repeats composed of lysine, alanine and proline function in binding nucleosome linker DNA.	histone H1
		• F/G	Disordered domains with phenylalanine-glycine repeats influence NPC gating behavior.	Nucleoporins
		• Etc.		
Protein interactions	Fuzzy complexes by topology [202]	• Polymorphic	A form of static disorder, with alternative bound conformations serving distinct functions by having different effects on the binding partner.	β-catenin ~ Tcf4, NLS ~ importin-α, actin ~ WH2 domain
		• Clamp	Complex formation through folding upon binding of two disordered protein segments, connected by a linker that remains disordered.	Ste5 ~ Fus3, Myosin VI ~ actin filament, Oct-1 ~ DNA
		• Flanking	Complex formation through folding upon binding of a central disordered protein segment, flanked by two regions that remain disordered.	SF1 splicing factor ~ U2AF, proline-rich peptides ~ SH3 domains, $p27^{Kip1}$ ~ cyclin-Cdk2
		• Random	Disordered regions that remain highly dynamic even in the bound state.	elastin self-assembly, Sic1 ~ Cdc4

	Fuzzy complexes by mechanism [204,211]	• Conformational selection	The fuzzy region facilitates the formation of the binding competent form by shifting the conformational equilibrium.	Max ~ DNA, MeCP2 ~ DNA
		• Flexibility modulation	The fuzzy region modulates the flexibility of the binding interface and changes binding entropy.	Ets-1 ~ DNA, SSB ~ DNA
		• Competitive binding	The fuzzy region serves as an intramolecular competitive partner for the binding surface.	HMGB1 ~ DNA, RNase1 ~ RNase inhibitor
		• Tethering	The fuzzy region increases the local concentration of a weak-affinity binding domain near the target, or anchors it via transient interactions.	RPA ~ DNA, UPF1 ~ UPF2, PC4 ~ VP16
	Binding plasticity [216]	• Static	Mono-/polyvalent complexes, chameleons, penetrators, huggers.	
		• Coiled-coil based	Intertwined strings, long cylindrical containers, connectors, armature, tweezers and forceps, grabbers, tentacles, pullers, stackers.	
		• Dynamic	Cloud contacts and protein interaction ensembles.	
Evolution	Sequence conservation [53]	• Flexible	Regions that require the property of disorder for functionality regardless of the exact sequence.	signaling and regulatory proteins (Sky1, Bur1)
		• Constrained	Regions of conserved disorder with also highly conserved amino acid sequences.	ribosomal proteins (Rpl5), protein chaperones (Hsp90)
		• Non-conserved	No conservation of the disorder, nor of the underlying sequence; no clear functional hallmarks.	yeast Ty1 retrotransposon domains A and B
	Conservation of amino acid composition [219]	• HR	IDRs with high residue conservation.	transcription regulation and DNA binding
		• LRHT	IDRs with low residue conservation but high conservation of the amino acid composition of the region.	ATPase and nuclease activities
		• LRLT	IDRs with neither conservation of sequence nor conservation of amino acid composition.	(metal) ion binding proteins
	Lineage and species specificity [137]	• Prokaryotes	Species from different kingdoms of life seem to use disorder for different types of functions.	longer lasting interactions involved in complex formation
		• Eukaryotes and viruses		transient interactions in signaling and regulation
	History and mechanism of repeat evolution [60]	• Type I	Repeats that showed no function diversification after expansion.	titin PEVK domain, salivary proline-rich proteins
		• Type II	Repeats that acquired diverse functions through mutation or differential location within the sequence.	RNA polymerase II (CTD)
		• Type III	Repeats that gained new functions as a consequence of their expansion.	prion protein octarepeats
Regulation	Expression patterns [173]	• Constitutive	IDPs encoded by constitutively highly expressed transcripts are almost entirely disordered and often ribosomal proteins.	ribosomal L proteins
		• High	IDP encoding transcripts showing high expression levels in most tissues and little tissue	protease inhibitors, splicing factors, complex assemblers

			specificity.	
		• Medium	These IDP encoding transcripts are expressed at medium levels, with some tissue-specificity.	DNA binding, transcription regulation
		• Tissue-specific	IDP encoding transcripts with highly tissue-specific expression.	cell organization regulators, complex disassemblers
		• Low or transient	IDP encoding transcripts that are present in undetectable amounts; more than half of analyzed IDPs.	variety of functions
	Alternative splicing [246,247,251,254,255]		Regulation and evolutionary patterns of inclusion and exclusion of IDR-encoding exons can give insights into whether the encoded IDR functions in protein regulation and interactions.	a tissue-specific peptide with a phosphosite in the TJP1 protein in mouse, a Mammalian-specific peptide in the PTB1 splicing regulator
	Degradation kinetics [256-258,260,262,263]	• Degradation accelerators	IDRs that can influence and accelerate proteasomal degradation of the protein they are part of.	
		• Others	IDRs that have no influence on protein half-life or increase it, e.g. because of sequence composition that is incompatible with the proteasomal mechanism of degradation.	
Biophysical properties	Phase transition [116,278]		Certain IDRs (such as those that contain specific low-complexity regions or interaction motifs) can undergo phase transitions like the formation of protein-based droplets or hydrogels.	Multivalent SH3-binding motifs in phase separation, granule-like assemblies of RNA-binding proteins containing low-complexity IDRs

Figure legends

Figure 1: Structured domains and Intrinsically Disordered Regions (IDRs) are two fundamental classes of functional building blocks of proteins. The synergy between disordered regions and structured domains increases the functional versatility of proteins.

Figure 2: The number of protein-coding genes in the human genome with various amounts of disorder. Histograms of the numbers of human genes with annotation (**A**) and without annotation (**B**), grouped by the percentage of disordered residues. (**C**) A comparison of the fraction of annotated and unannotated human genes with different amounts of disorder. Residues in each protein are defined as disordered when there is a consensus between >75% of the predictors in the D^2P^2 database[48] at that position. The set of human genes was taken from ENSEMBL release 63,[1] and the representative protein coded for by the longest transcript was used in each case. The annotation was taken from the description field with ‘open reading frame’, ‘hypothetical’, ‘uncharacterized’ and ‘putative protein’ treated as no annotation.

Figure 3: The fraction of disordered residues located in domains in human protein-coding genes: (**A**) residues inside (*left*) and outside (*right*) of SCOP domains,[24] and (**B**) residues inside (*left*) and outside (*right*) of Pfam domains.[22] The SCOP domains in human proteins are defined by the SUPERFAMILY database.[10] Disordered residues were taken from the D^2P^2 database[48] (when there is a consensus between >75% of the disorder predictors). The set of human genes was taken from ENSEMBL release 63.[1]

Figure 4: Functional classification scheme of IDRs. The function of disordered regions can stem directly from their highly flexible nature, when they fulfill entropic chain functions (such as linkers and spacers, indicated in dark-tone red), or from their ability to bind to partner molecules (proteins, other macromolecules, or small molecules). In the latter case, they bind either transiently as display sites of post-translational modifications or as chaperones (indicated in green), or they bind permanently as effectors, assemblers or scavengers (indicated in dark-tone blue). More extensive descriptions and examples are found in the main text. Adapted from Tompa (2005)[57].

Figure 5: Functional classification of IDRs according to their interaction features. (**A**) The flexibility of IDRs facilitates access to enzymes that catalyze post-translational modifications and effectors that bind these PTMs. This permits combinatorial regulation and re-use of the same components in multiple biological processes. (**B**) The availability of molecular recognition features and linear motifs within the IDRs enable the fishing for (‘fly-casting’) and gathering of different partners. (**C**) Conformational variability enables a nearly perfect molding to fit the binding interfaces of very diverse interaction partners. Context-dependent folding of an IDR can activate signaling processes in one case or inhibit them in another, resulting in completely different outcomes.

Figure 6: Functional classification of linear motifs. Linear motifs can be divided into two major families, which each have three further subgroups. The modification class motifs all act as recognition sites for enzyme active sites, whereas the ligand class motifs are always recognized by the binding surface of a protein partner. More detailed classification beyond the graph shown here is possible. For example, an important subgroup of docking motifs are the degrons, which regulate protein stability by recruiting members of the ubiquitin-proteasome system. In the regular expressions, x corresponds to any amino acid, while other letters represent single letter codes of amino acids; letters within square brackets mean either residue is allowed in that position.

Figure 7: Classification of molecular recognition features (MoRFs) based on the secondary structure of the bound state. MoRFs (red ribbons) undergo disorder-to-order transition upon binding their partners (blue surfaces). (**A**) α-MoRF. BH3 domain of BAD (MoRF) bound to bcl-xl (partner) (PDB ID: 1G5J). (**B**) β-MoRF. Inhibitor of apoptosis protein DIAP1 (partner) bound to N-terminus of cell death protein GRIM (MoRF) (PDB ID: 1JD5). (**C**) ι -MoRF. AP-2 (partner) bound to the recognition motif of amphiphysin (MoRF) (PDB ID: 1KY7). (**D**) complex-MoRF. Phosphotyrosine-binding domain (PTB) of the X11 protein (partner) bound to amyloid beta A4 protein (MoRF) (PDB ID: 1X11). Note that the PTB domain of X11 actually binds unphosphorylated peptides and is a PTB by sequence similarity. Reproduced from Vacic et al. (2007)[102]. (**E**) Promiscuity of disorder-controlled interactions illustrated by the p53 interaction network. A structure versus disorder prediction on the p53 amino acid sequence is shown in the center of the figure (up = disorder, down = order) along with the structures of various regions of p53 bound to fourteen different partners. The predictions for a

central region of structure and disordered amino and carbonyl termini have been confirmed experimentally for p53. The various regions of p53 are color coded to show their structures in the complex and to map the binding segments to the amino acid sequence. Starting with the p53-DNA complex (top, left, magenta protein, blue DNA), and moving in a clockwise direction, the Protein Data Bank[125] IDs and partner names are given as follows for the fourteen complexes: (1tsr – DNA), (1gzh – 53BP1), (1q2d – gcn5), (3sak – p53 (tetramerization domain)), (1xqh – set9), (1h26 – cyclinA), (1ma3 – sirtuin), (1jsp – CBP bromo domain), (1dt7 – s100ββ), (2h1l – sv40 Large T antigen), (1ycs – 53BP2), (2gs0 – PH), (1ycr – MDM2), and (2b3g – RPA70). Reproduced from Uversky and Dunker (2010)[39].

Figure 8: Schematic representation of the continuum model of protein structure. The color gradient represents a continuum of conformational states ranging from highly dynamic, expanded conformational ensembles (red) to compact, dynamically restricted, fully folded globular states (blue). Dynamically disordered states are represented by heavy lines, stably folded structures as cartoons. A characteristic of IDPs is that they rapidly interconvert between multiple states in the dynamic conformational ensemble. In the continuum model, the proteome would populate the entire spectrum of dynamics, disorder, and folded structure depicted.

Figure 9: The protein quartet model of protein conformational states. In accordance with this model, protein function arises from four types of conformations of the polypeptide chain (ordered forms, molten globules, pre-molten globules, and random coils) and transitions between any of these states.

Figure 10: Original[144] and modified[167] diagram-of-states to classify predicted conformational properties of IDPs (and IDRs modeled as IDPs). (**A**) The original diagram predicts that sequences with a net charge per residue above 0.25 will be swollen coils. The three axes denote the fraction of positively charged residues, f_+, the fraction of negatively charged residues, f_-, and the hydropathy. All three parameters are calculated from the amino acid composition. Green dots correspond to 364 curated disordered sequences extracted from the DisProt database.[166] These sequences have hydropathy values that designate them as being disordered, i.e. they lie in the bottom portion of the pyramid by definition. Additional filters were used for chain length (more than 30 residues) and the fraction of proline residues ($f_{pro} < 0.3$). 97% of sequences used in this annotation have a net charge per residue of less than 0.26 and are thus predicted to be globule formers.[167] (**B**) Modified diagram-of-states from panel (A) with a focus only on the bottom portion of the pyramid (i.e. stipulating that the hydropathy is low enough to be ignored).[167] The polyampholytic contribution expands the space encompassed by non-globule-formers by subdividing the disordered globules space in panel (A) into three distinct regions of which sequences in regions 2 and 3 actually may not form globules. In these polyampholytic regions, one has to account for the total charge, in terms of the fraction of charged residues (FCR), as well as the net charge per residue (NCPR) as opposed to NCPR alone. Conformations in regions 2 and 3 are expected to be random-coil-like if oppositely charged residues are well mixed in the linear sequence. Otherwise, one can expect compact or semi-compact conformations. The classification scheme uses only the amino acid sequence as input.

Figure 11: Classification of fuzzy complexes by topology (upper panel) and by mechanism (lower panel). Blue arrows indicate interactions between fuzzy disordered regions and structured molecules. Protein Data Bank[125] identifiers for the structures are given in parentheses. **Topological categories:** (**A**) Polymorphic. The WH2 domain of cibulot interacts with actin in alternative locations: via an 18-residue segment (3u9z) or via only three residues (2ff3). The flanking regions remain dynamically disordered. (**B**) Clamp. The Oct-1 transcription factor has a bipartite DNA recognition motif. The two globular binding domains are connected by a 23 residue long disordered linker (1hf0), shortening of which reduces binding affinity. (**C**) Flanking. The $p27^{Kip1}$ cell-cycle kinase inhibitor binds to the cyclin-Cdk2 complex (1jsu). The kinase binding site is flanked by a ~100 residue long disordered linker, which enables T187 at the C-terminus to be phosphorylated. (**D**) Random. UmuD2 is a dimer that is produced from UmuD by RecA-facilitated self-cleavage (1i4v). The resulting proteins exhibit a random coil signal in circular dichroism experiments at physiologically relevant concentrations. **Mechanistic categories:** (**E**) Conformational selection. The fuzzy N-terminal acidic tail of the Max transcription factor (1nkp) facilitates formation of the DNA binding helix (dark red) of the leucine zipper basic helix-loop-helix (bHLH) motif. (**F**) Flexibility modulation. The disordered serine/arginine-rich region of Ets-1 transcription factor (1mdm) changes DNA binding affinity by 100–1000-fold by modulating the flexibility of the binding segment via transient interactions. (**G**) Competitive binding. The acidic fuzzy C-terminal tail of high-mobility group protein B1 (2gzk) competes with DNA for the positively charged binding surfaces. (**H**) Tethering. The binding of the virion protein 16 activation domain to the human transcriptional coactivator positive cofactor 4 (2phe) is facilitated by acidic disordered regions, which anchor the binding segments.

Figure 12: A portrait gallery of disorder-based complexes. Illustrative examples of various interaction modes of intrinsically disordered proteins are shown. Protein Data Bank[125] identifiers for the structures are given in parentheses. (**A**) MoRFs. **Aa**, α-MoRF, a complex between the botulinum neurotoxin (red helix) and its receptor (a blue cloud) (2NM1); **Ab**, ι-MoRF, a complex between an 18-mer cognate peptide derived from the α1 subunit of the nicotinic acetylcholine receptor from *Torpedo californica* (red helix) and α-cobratoxin (a blue cloud) (1LXH). (**B**) Wrappers. **Ba**, rat PP1 (blue cloud) complexed with mouse inhibitor-2 (red helices) (2O8A); **Bb**, a complex between the paired domain from the Drosophila paired (prd) protein and DNA (1PDN). (**C**). Penetrator. Ribosomal protein s12 embedded into the rRNA (1N34). (**D**). Huggers. **Da**, *E. coli trp* repressor dimer (1ZT9); **Db**, tetramerization domain of p53 (1PES); **Dc**, tetramerization domain of p73 (2WQI). (**E**) Intertwined strings. **Ea**, dimeric coiled coil, a basic coiled-coil protein from *Eubacterium eligens* ATCC 27750 (3HNW); **Eb**, trimeric coiled coil, salmonella trimeric autotransporter adhesin, SadA (2WPQ); **Ec**, tetrameric coiled coil, the virion-associated protein P3 from Caulimovirus (2O1J). (**F**) Long cylindrical containers. **Fa**, pentameric coiled coil, side and top views of the assembly domain of cartilage oligomeric matrix protein (1FBM); **Fb**, side and top views of the seven-helix coiled coil, engineered version of the GCN4 leucine zipper (2HY6). (**G**) Connectors. **Ga**, human heat shock factor binding protein 1 (3CI9); **Gb**, the bacterial cell division protein ZapA from *Pseudomonas aeruginosa* (1W2E). (**H**) Armature. **Ha**, side and top views of the envelope glycoprotein GP2 from Ebola virus (2EBO); **Hb**, side and top views of a complex between the N- and C-terminal peptides derived from the membrane fusion protein of the Visna (1JEK). (**I**) Tweezers or forceps. A complex between c-Jun, c-Fos and DNA. Proteins are shown as red helices, whereas DNA is shown as a blue cloud (1FOS). (**J**) Grabbers. Structure of the complex between βPIX coiled coil (red helices) and Shank PDZ (blue cloud) (3L4F). (**K**) Tentacles. Structure of the hexameric molecular chaperone prefoldin from the archaeum *Methanobacterium thermoautotrophicum* (1FXK). (**L**) Pullers. Structure of the ClpB chaperone from *Thermus thermophilus* (1QVR). (**M**) Chameleons. The C-terminal fragment of p53 gains different types of secondary structure in complexes with four different binding partners, cyclinA (1H26), sirtuin (1MA3), CBP bromo domain (1JSP), and s100ββ (1DT7). (**N**) Stackers or β-arcs. **Na**, stack of β-arches, β-amyloid; **Nb**, superpleated β-structure (Sup35p, Ure2P, α-synuclein); **Nc**, stack of β-solenoids (prion); **Nd**, stack of β-arch dimers (insulin); **Ne**, β-solenoids. Modified from Kajava et al. (2010)[311]. (**O**) Dynamic complexes. Schematic representation of the polyelectrostatic model of Sic1-Cdc4 interaction. Schematic representation of an IDP (ribbon) interacting with a folded receptor (grey shape) through several distinct binding motifs and an ensemble of conformations (indicated by four representations of the interaction). The intrinsically disordered protein possesses positive and negative charges (depicted as blue and red circles, respectively) giving rise to a net charge q_l, while the binding site in the receptor (light blue) has a charge q_r. The effective distance $<r>$ is between the binding site and the center of mass of the intrinsically disordered protein. Reproduced from Mittag et al. (2010)[203].

Figure 13: Classification of disordered regions according to their evolutionary conservation (constrained, flexible and non-conserved disorder). (**A**) Schematic of computing disorder conservation and amino acid sequence conservation. The alignments are used to calculate the percentage of sequences in which a residue is disordered and the percentage of sequences in which the amino acid itself is conserved. A residue is considered to be conserved disordered if the property of disorder is conserved in at least half of species. Similarly, the amino acid type of a residue is considered conserved if it is present in at least half of species. Disordered residues in which both sequence and disorder are conserved are referred to as constrained disorder. Disordered residues in which disorder is conserved but not the amino acid sequence are referred to as flexible disorder. Residues which are disordered in *S. cerevisiae* but not cases of conserved disorder are referred to as non-conserved disorder. (**B**) Disorder splits into three distinct phenomena. Functional enrichment maps of proteins enriched in flexible disorder versus constrained disorder. The area of each rectangle is proportional to the occurrence of that type of disorder in the alignments. Related gene ontology terms are grouped based on gene overlap. Adapted from Bellay et al. (2011)[53].

Figure 14: Repeat expansion creates IDRs. IDRs are abundant in repeating sequence elements, which suggest that repeat expansion is an important mechanism by which genetic material encoding for structural disorder is generated. The expanding repeats may fall into three classes (types) in terms of their functional diversification following expansion. Individual repeats may remain functionally equivalent (type I), or diversify (type II), or collectively acquire a completely new function (type III). Dark-tone red indicates structural disorder of the repeat, which may undergo full (dark-tone blue) or partial (green) induced folding upon binding to a partner. Adapted from Tompa (2003)[60].

Figure 15: A summary of expression–function trends for human transcripts encoding highly disordered proteins. The x-axis represents the ($\log_{10}$) number of tissues in which the transcript is expressed; the y-axis represents the $\log_{10}$ average magnitude of expression within the tissues. From the data, five distinct functional classes of highly disordered human proteins become apparent.

Figure 16: Transcriptional and post-transcriptional gene regulation can be informative of IDR function. How inclusion of exons that code for IDRs is regulated during gene transcription and alternative splicing can give insights into the functional roles of the encoded disordered regions. For example, tissue- or developmental-specific regulation of alternative splicing or alternative promoter and polyadenylation site usage can be associated with important roles of the encoded IDRs in protein regulation and cellular interactions through for example the presence of binding motifs and phosphosites. Additionally, information on the conservation of patterns of exon inclusion (i.e. events shared among different evolutionary lineages versus species-specific events) can aid in better characterization of the encoded IDRs. The figure illustrates a hypothetical example where an exon (largest red box) that is included in a tissue-specific manner both in human and mouse encodes an IDR that embeds a phosphosite (P) and is involved in protein regulation. The human gene depicted in the figure has an additional exon (smallest red box), which encodes an IDR with a short interaction motif and which is also included in a tissue-specific manner in humans. Gene structures, mature mRNAs and corresponding protein isoforms are shown for human and mouse brain and heart tissues. On the right, possible functional roles of the IDRs encoded by the brain isoforms are illustrated. The examples illustrate how protein functional space can increase due to alternative splicing of exons that encode IDRs.

Figure 17: Involvement of IDRs in phase transitions. (**A**) Interactions between proteins that contain multiple copies of a specific domain (an SH3 domain in the figure) and IDRs with multiple instances of its interaction motif (proline-rich SH3 motif here) can, at appropriate concentrations, produce sharp liquid–liquid-demixing phase separations. This phase transition is likely to increase local 'active' protein concentrations exploitable for signaling switches. (**B**) High concentrations of low-complexity IDRs found in certain RNA binding domains lead to a reversible phase transition with the formation of highly dynamic hydrogels. These RNA granule-like assemblies consist of heteromeric protein aggregates and allow localization and storage of functionally related but non-identical RNA molecules.

Figure Box 3: Disordered residues generally evolve more quickly than structured residues, but do display a wide range of evolutionary rates. Boxplots of the distributions of evolutionary rates for the predicted structured (blue) and disordered (red) residues across the human proteome. A residue with a high evolutionary rate is less conserved. Evolutionary rates were calculated using the rate4site algorithm[309] on homolog alignments from 56 primarily vertebrate species.[1] Disordered residues were predicted using MULTICOM-REFINE.[308] Filled boxes represent the 50% of data points in the two quartiles above and below the median (which is represented by the horizontal bar within each box). Vertical lines (whiskers) connected to the boxes represent the highest and lowest non-outlier data points, with outliers being defined by a distance of more than 1.5 times the interquartile range from the median. Outliers are not shown for visual clarity. The distributions of evolutionary rates for disordered and structured residues differ significantly (P value < 10^{-15}, Mann–Whitney U test).

References

(1) Flicek, P.; Ahmed, I.; Amode, M. R.; Barrell, D.; Beal, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fairley, S.; et al. *Nucleic Acids Res.* **2013**, *41*, D48.

(2) NCBI Resource Coordinators *Nucleic Acids Res.* **2013**, *41*, D8.

(3) Kolodny, R.; Pereyaslavets, L.; Samson, A. O.; Levitt, M. *Annu. Rev. Biophys.* **2013**, *42*, 559.

(4) Raes, J.; Harrington, E. D.; Singh, A. H.; Bork, P. *Curr. Opin. Struct. Biol.* **2007**, *17*, 362.

(5) Jaroszewski, L.; Li, Z.; Krishna, S. S.; Bakolitsa, C.; Wooley, J.; Deacon, A. M.; Wilson, I. A.; Godzik, A. *PLoS Biol.* **2009**, *7*, e1000205.

(6) Levitt, M. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 11079.

(7) The UniProt Consortium *Nucleic Acids Res.* **2012**, *40*, D71.

(8) Aravind, L.; Koonin, E. V. *J. Mol. Biol.* **1999**, *287*, 1023.

(9) Gough, J.; Karplus, K.; Hughey, R.; Chothia, C. *J. Mol. Biol.* **2001**, *313*, 903.

(10) de Lima Morais, D. A.; Fang, H.; Rackham, O. J.; Wilson, D.; Pethica, R.; Chothia, C.; Gough, J. *Nucleic Acids Res.* **2011**, *39*, D427.

(11) Aravind, L. *Genome Res.* **2000**, *10*, 1074.

(12) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. *Nat. Genet.* **2000**, *25*, 25.

(13) Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; Yeates, T. O. *Nature* **2000**, *405*, 823.

(14) Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; Orengo, C. A. *Nat. Struct. Biol.* **2000**, *7 Suppl*, 991.

(15) Whisstock, J. C.; Lesk, A. M. *Q. Rev. Biophys.* **2003**, *36*, 307.

(16) Gabaldon, T.; Huynen, M. A. *Cell. Mol. Life Sci.* **2004**, *61*, 930.

(17) Frishman, D. *Chem. Rev.* **2007**, *107*, 3448.

(18) Chothia, C.; Lesk, A. M. *EMBO J.* **1986**, *5*, 823.

(19) Laskowski, R. A.; Thornton, J. M. *Nat. Rev. Genet.* **2008**, *9*, 141.

(20) Kim, S. H.; Shin, D. H.; Choi, I. G.; Schulze-Gahmen, U.; Chen, S.; Kim, R. *J. Struct. Funct. Genomics* **2003**, *4*, 129.

(21) Dessailly, B. H.; Nair, R.; Jaroszewski, L.; Fajardo, J. E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. *Structure* **2009**, *17*, 869.

(22) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; et al. *Nucleic Acids Res.* **2012**, *40*, D290.

(23) Bateman, A.; Coggill, P.; Finn, R. D. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2010**, *66*, 1148.

(24) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536.

(25) Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. *Nucleic Acids Res.* **2008**, *36*, D419.

(26) Sillitoe, I.; Cuff, A. L.; Dessailly, B. H.; Dawson, N. L.; Furnham, N.; Lee, D.; Lees, J. G.; Lewis, T. E.; Studer, R. A.; Rentzsch, R.; et al. *Nucleic Acids Res.* **2013**, *41*, D490.

(27) Lewis, T. E.; Sillitoe, I.; Andreeva, A.; Blundell, T. L.; Buchan, D. W.; Chothia, C.; Cuff, A.; Dana, J. M.; Filippis, I.; Gough, J.; et al. *Nucleic Acids Res.* **2013**, *41*, D499.

(28) Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 11504.

(29) Daughdrill, G. W.; Chadsey, M. S.; Karlinsey, J. E.; Hughes, K. T.; Dahlquist, F. W. *Nat. Struct. Biol.* **1997**, *4*, 285.

(30) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321.

(31) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. *Proteins* **2000**, *41*, 415.

(32) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; et al. *J. Mol. Graph. Model.* **2001**, *19*, 26.

(33) Tompa, P. *Trends Biochem. Sci.* **2002**, *27*, 527.

(34) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739.

(35) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. *J. Mol. Biol.* **2004**, *337*, 635.

(36) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197.

(37) Dunker, A. K.; Oldfield, C. J.; Meng, J.; Romero, P.; Yang, J. Y.; Chen, J. W.; Vacic, V.; Obradovic, Z.; Uversky, V. N. *BMC Genomics* **2008**, *9 Suppl 2*, S1.

(38) Gsponer, J.; Babu, M. M. *Prog. Biophys. Mol. Biol.* **2009**, *99*, 94.

(39) Uversky, V. N.; Dunker, A. K. *Biochim. Biophys. Acta* **2010**, *1804*, 1231.

(40) Tompa, P. *Trends Biochem. Sci.* **2012**, *37*, 509.

(41) Forman-Kay, J. D.; Mittag, T. *Structure* **2013**, *21*, 1492.

(42) Dunker, A. K.; Babu, M. M.; Barbar, E.; Blackledge, M.; Bondos, S. E.; Dosztányi, Z.; Dyson, H. J.; Forman-Kay, J.; Fuxreiter, M.; Gsponer, J.; et al. *Intrinsically Disordered Proteins* **2013**, *1*, e24157.

(43) Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. *Proteins* **2001**, *42*, 38.

(44) Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; Dunker, A. K. *Nucleic Acids Res.* **2004**, *32*, 1037.

(45) Collins, M. O.; Yu, L.; Campuzano, I.; Grant, S. G.; Choudhary, J. S. *Mol. Cell. Proteomics* **2008**, *7*, 1331.

(46) Diella, F.; Haslam, N.; Chica, C.; Budd, A.; Michael, S.; Brown, N. P.; Trave, G.; Gibson, T. J. *Front. Biosci.* **2008**, *13*, 6580.

(47) Davey, N. E.; Van Roey, K.; Weatheritt, R. J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T. J. *Mol. BioSyst.* **2012**, *8*, 268.

(48) Oates, M. E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M. J.; Xue, B.; Dosztanyi, Z.; Uversky, V. N.; Obradovic, Z.; Kurgan, L.; et al. *Nucleic Acids Res.* **2013**, *41*, D508.

(49) Babu, M. M.; Kriwacki, R. W.; Pappu, R. V. *Science* **2012**, *337*, 1460.

(50) Brown, C. J.; Takayama, S.; Campen, A. M.; Vise, P.; Marshall, T. W.; Oldfield, C. J.; Williams, C. J.; Dunker, A. K. *J. Mol. Evol.* **2002**, *55*, 104.

(51) Chen, J. W.; Romero, P.; Uversky, V. N.; Dunker, A. K. *J. Proteome Res.* **2006**, *5*, 879.

(52) Brown, C. J.; Johnson, A. K.; Daughdrill, G. W. *Mol. Biol. Evol.* **2010**, *27*, 609.

(53) Bellay, J.; Han, S.; Michaut, M.; Kim, T.; Costanzo, M.; Andrews, B. J.; Boone, C.; Bader, G. D.; Myers, C. L.; Kim, P. M. *Genome Biol.* **2011**, *12*, R14.

(54) Brown, C. J.; Johnson, A. K.; Dunker, A. K.; Daughdrill, G. W. *Curr. Opin. Struct. Biol.* **2011**, *21*, 441.

(55) Nilsson, J.; Grahn, M.; Wright, A. P. *Genome Biol.* **2011**, *12*, R65.

(56) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573.

(57) Tompa, P. *FEBS Lett.* **2005**, *579*, 3346.

(58) Uversky, V. N. *FEBS Lett.* **2013**, *587*, 1891.

(59) Daughdrill, G. W.; Narayanaswami, P.; Gilmore, S. H.; Belczyk, A.; Brown, C. J. *J. Mol. Evol.* **2007**, *65*, 277.

(60) Tompa, P. *BioEssays* **2003**, *25*, 847.

(61) Tskhovrebova, L.; Trinick, J. *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 679.

(62) Seet, B. T.; Dikic, I.; Zhou, M. M.; Pawson, T. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 473.

(63) Vucetic, S.; Xie, H.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Obradovic, Z.; Uversky, V. N. *J. Proteome Res.* **2007**, *6*, 1899.

(64) Uversky, V. N. *Curr. Pharm. Des.* **2013**, *19*, 4191.

(65) Galea, C. A.; Wang, Y.; Sivakolundu, S. G.; Kriwacki, R. W. *Biochemistry* **2008**, *47*, 7598.

(66) Holt, L. J.; Tuch, B. B.; Villen, J.; Johnson, A. D.; Gygi, S. P.; Morgan, D. O. *Science* **2009**, *325*, 1682.

(67) Gsponer, J.; Futschik, M. E.; Teichmann, S. A.; Babu, M. M. *Science* **2008**, *322*, 1365.

(68) Landry, C. R.; Levy, E. D.; Michnick, S. W. *Trends Genet.* **2009**, *25*, 193.

(69) Van Roey, K.; Gibson, T. J.; Davey, N. E. *Curr. Opin. Struct. Biol.* **2012**, *22*, 378.

(70) Van Roey, K.; Dinkel, H.; Weatheritt, R. J.; Gibson, T. J.; Davey, N. E. *Sci. Signal.* **2013**, *6*, rs7.

(71) Fuxreiter, M.; Tompa, P.; Simon, I. *Bioinformatics* **2007**, *23*, 950.

(72) Gibson, T. J. *Trends Biochem. Sci.* **2009**, *34*, 471.

(73) Perkins, J. R.; Diboun, I.; Dessailly, B. H.; Lees, J. G.; Orengo, C. *Structure* **2010**, *18*, 1233.

(74) Kouzarides, T. *Cell* **2007**, *128*, 693.

(75) Galea, C. A.; Nourse, A.; Wang, Y.; Sivakolundu, S. G.; Heller, W. T.; Kriwacki, R. W. *J. Mol. Biol.* **2008**, *376*, 827.

(76) Bode, A. M.; Dong, Z. *Nat. Rev. Cancer* **2004**, *4*, 793.

(77) Schroeder, R.; Barta, A.; Semrad, K. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 908.

(78) Young, J. C.; Agashe, V. R.; Siegers, K.; Hartl, F. U. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 781.

(79) Tompa, P.; Csermely, P. *FASEB J.* **2004**, *18*, 1169.

(80) Ivanyi-Nagy, R.; Davidovic, L.; Khandjian, E. W.; Darlix, J. L. *Cell. Mol. Life Sci.* **2005**, *62*, 1409.

(81) Kovacs, D.; Tompa, P. *Biochem. Soc. Trans.* **2012**, *40*, 963.

(82) Reichmann, D.; Xu, Y.; Cremers, C. M.; Ilbert, M.; Mittelman, R.; Fitzgerald, M. C.; Jakob, U. *Cell* **2012**, *148*, 947.

(83) Sugase, K.; Dyson, H. J.; Wright, P. E. *Nature* **2007**, *447*, 1021.

(84) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31.

(85) Mucsi, Z.; Hudecz, F.; Hollosi, M.; Tompa, P.; Friedrich, P. *Protein Sci.* **2003**, *12*, 2327.

(86) Kim, A. S.; Kakalis, L. T.; Abdul-Manan, N.; Liu, G. A.; Rosen, M. K. *Nature* **2000**, *404*, 151.

(87) Trudeau, T.; Nassar, R.; Cumberworth, A.; Wong, E. T.; Woollard, G.; Gsponer, J. *Structure* **2013**, *21*, 332.

(88) Peng, Z.; Oldfield, C. J.; Xue, B.; Mizianty, M. J.; Dunker, A. K.; Kurgan, L.; Uversky, V. N. *Cell. Mol. Life Sci.* **2013**.

(89) Fuxreiter, M.; Tompa, P.; Simon, I.; Uversky, V. N.; Hansen, J. C.; Asturias, F. J. *Nat. Chem. Biol.* **2008**, *4*, 728.

(90) Hegyi, H.; Schad, E.; Tompa, P. *BMC Struct. Biol.* **2007**, *7*, 65.

(91) Haynes, C.; Oldfield, C. J.; Ji, F.; Klitgord, N.; Cusick, M. E.; Radivojac, P.; Uversky, V. N.; Vidal, M.; Iakoucheva, L. M. *PLoS Comput. Biol.* **2006**, *2*, e100.

(92) Kim, P. M.; Sboner, A.; Xia, Y.; Gerstein, M. *Mol. Syst. Biol.* **2008**, *4*, 179.
(93) Gunasekaran, K.; Tsai, C. J.; Kumar, S.; Zanuy, D.; Nussinov, R. *Trends Biochem. Sci.* **2003**, *28*, 81.
(94) Xue, B.; Romero, P. R.; Noutsou, M.; Maurice, M. M.; Rudiger, S. G.; William, A. M., Jr.; Mizianty, M. J.; Kurgan, L.; Uversky, V. N.; Dunker, A. K. *FEBS Lett.* **2013**, *587*, 1587.
(95) Cortese, M. S.; Uversky, V. N.; Dunker, A. K. *Prog. Biophys. Mol. Biol.* **2008**, *98*, 85.
(96) Buday, L.; Tompa, P. *FEBS J.* **2010**, *277*, 4348.
(97) Daniels, A. J.; Williams, R. J.; Wright, P. E. *Neuroscience* **1978**, *3*, 573.
(98) Ren, S.; Uversky, V. N.; Chen, Z.; Dunker, A. K.; Obradovic, Z. *BMC Genomics* **2008**, *9 Suppl 2*, S26.
(99) Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. *J. Mol. Biol.* **2004**, *338*, 1015.
(100) Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Romero, P.; Uversky, V. N.; Dunker, A. K. *Biochemistry* **2005**, *44*, 12454.
(101) Mohan, A.; Oldfield, C. J.; Radivojac, P.; Vacic, V.; Cortese, M. S.; Dunker, A. K.; Uversky, V. N. *J. Mol. Biol.* **2006**, *362*, 1043.
(102) Vacic, V.; Oldfield, C. J.; Mohan, A.; Radivojac, P.; Cortese, M. S.; Uversky, V. N.; Dunker, A. K. *J. Proteome Res.* **2007**, *6*, 2351.
(103) Lee, S. H.; Kim, D. H.; Han, J. J.; Cha, E. J.; Lim, J. E.; Cho, Y. J.; Lee, C.; Han, K. H. *Curr. Protein Pept. Sci.* **2012**, *13*, 34.
(104) Hinds, M. G.; Smits, C.; Fredericks-Short, R.; Risk, J. M.; Bailey, M.; Huang, D. C.; Day, C. L. *Cell Death Differ.* **2007**, *14*, 128.
(105) Dinkel, H.; Michael, S.; Weatheritt, R. J.; Davey, N. E.; Van Roey, K.; Altenberg, B.; Toedt, G.; Uyar, B.; Seiler, M.; Budd, A.; et al. *Nucleic Acids Res.* **2012**, *40*, D242.
(106) Mi, T.; Merlin, J. C.; Deverasetty, S.; Gryk, M. R.; Bill, T. J.; Brooks, A. W.; Lee, L. Y.; Rathnayake, V.; Ross, C. A.; Sargeant, D. P.; et al. *Nucleic Acids Res.* **2012**, *40*, D252.
(107) Stein, A.; Ceol, A.; Aloy, P. *Nucleic Acids Res.* **2011**, *39*, D718.
(108) Weatheritt, R. J.; Luck, K.; Petsalaki, E.; Davey, N. E.; Gibson, T. J. *Bioinformatics* **2012**, *28*, 976.
(109) Davey, N. E.; Trave, G.; Gibson, T. J. *Trends Biochem. Sci.* **2011**, *36*, 159.
(110) Jurgens, M. C.; Voros, J.; Rautureau, G. J.; Shepherd, D. A.; Pye, V. E.; Muldoon, J.; Johnson, C. M.; Ashcroft, A. E.; Freund, S. M.; Ferguson, N. *Nat. Chem. Biol.* **2013**.
(111) Pop, C.; Salvesen, G. S. *J. Biol. Chem.* **2009**, *284*, 21777.
(112) Fischer, U.; Janicke, R. U.; Schulze-Osthoff, K. *Cell Death Differ.* **2003**, *10*, 76.
(113) Pines, J. *Biochem. J.* **1995**, *308 (Pt 3)*, 697.
(114) Zhou, X. Z.; Kops, O.; Werner, A.; Lu, P. J.; Shen, M.; Stoller, G.; Kullertz, G.; Stark, M.; Fischer, G.; Lu, K. P. *Mol. Cell* **2000**, *6*, 873.
(115) Pawson, T.; Nash, P. *Science* **2003**, *300*, 445.
(116) Li, P.; Banjade, S.; Cheng, H. C.; Kim, S.; Chen, B.; Guo, L.; Llaguno, M.; Hollingsworth, J. V.; King, D. S.; Banani, S. F.; et al. *Nature* **2012**, *483*, 336.
(117) Calderon-Villalobos, L. I.; Tan, X.; Zheng, N.; Estelle, M. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a005546.
(118) Pfleger, C. M.; Kirschner, M. W. *Genes Dev.* **2000**, *14*, 655.
(119) He, J.; Chao, W. C.; Zhang, Z.; Yang, J.; Cronin, N.; Barford, D. *Mol. Cell* **2013**, *50*, 649.
(120) Kalderon, D.; Roberts, B. L.; Richardson, W. D.; Smith, A. E. *Cell* **1984**, *39*, 499.
(121) Evans, P. R.; Owen, D. J. *Curr. Opin. Struct. Biol.* **2002**, *12*, 814.
(122) Balagopalan, L.; Coussens, N. P.; Sherman, E.; Samelson, L. E.; Sommers, C. L. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a005512.
(123) Burgen, A. S.; Roberts, G. C.; Feeney, J. *Nature* **1975**, *253*, 753.
(124) Espinoza-Fonseca, L. M. *Biochem. Biophys. Res. Commun.* **2009**, *382*, 479.
(125) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlic, A.; Quesada, M.; et al. *Nucleic Acids Res.* **2013**, *41*, D475.
(126) Vise, P. D.; Baral, B.; Latos, A. J.; Daughdrill, G. W. *Nucleic Acids Res.* **2005**, *33*, 2061.
(127) Borcherds, W.; Kashtanov, S.; Wu, H.; Daughdrill, G. W. *Proteins* **2013**.
(128) Chi, S. W.; Lee, S. H.; Kim, D. H.; Ahn, M. J.; Kim, J. S.; Woo, J. Y.; Torizawa, T.; Kainosho, M.; Han, K. H. *J. Biol. Chem.* **2005**, *280*, 38795.
(129) Bochkareva, E.; Kaustov, L.; Ayed, A.; Yi, G. S.; Lu, Y.; Pineda-Lucena, A.; Liao, J. C.; Okorokov, A. L.; Milner, J.; Arrowsmith, C. H.; et al. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15412.
(130) Worrall, J. A.; Gorna, M.; Pei, X. Y.; Spring, D. R.; Nicholson, R. L.; Luisi, B. F. *Biochem. Soc. Trans.* **2007**, *35*, 502.
(131) Cheng, Y.; Oldfield, C. J.; Meng, J.; Romero, P.; Uversky, V. N.; Dunker, A. K. *Biochemistry* **2007**, *46*, 13468.
(132) Chandran, V.; Luisi, B. F. *J. Mol. Biol.* **2006**, *358*, 8.
(133) Nurmohamed, S.; Vaidialingam, B.; Callaghan, A. J.; Luisi, B. F. *J. Mol. Biol.* **2009**, *389*, 17.
(134) Das, R. K.; Crick, S. L.; Pappu, R. V. *J. Mol. Biol.* **2012**, *416*, 287.

(135) Das, R. K.; Mao, A. H.; Pappu, R. V. *Sci. Signal.* **2012**, *5*, pe17.
(136) Tompa, P.; Fuxreiter, M.; Oldfield, C. J.; Simon, I.; Dunker, A. K.; Uversky, V. N. *BioEssays* **2009**, *31*, 328.
(137) Chen, J. W.; Romero, P.; Uversky, V. N.; Dunker, A. K. *J. Proteome Res.* **2006**, *5*, 888.
(138) Buljan, M.; Frankish, A.; Bateman, A. *Genome Biol.* **2010**, *11*, R74.
(139) Pentony, M. M.; Jones, D. T. *Proteins* **2010**, *78*, 212.
(140) Teraguchi, S.; Patil, A.; Standley, D. M. *BMC Bioinformatics* **2010**, *11 Suppl 7*, S7.
(141) Meszaros, B.; Dosztanyi, Z.; Simon, I. *PLoS ONE* **2012**, *7*, e46829.
(142) Demarest, S. J.; Martinez-Yamout, M.; Chung, J.; Chen, H.; Xu, W.; Dyson, H. J.; Evans, R. M.; Wright, P. E. *Nature* **2002**, *415*, 549.
(143) Dyson, H. J.; Wright, P. E. *Nat. Struct. Biol.* **1998**, *5 Suppl*, 499.
(144) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 8183.
(145) Muller-Spath, S.; Soranno, A.; Hirschfeld, V.; Hofmann, H.; Ruegger, S.; Reymond, L.; Nettels, D.; Schuler, B. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 14609.
(146) Receveur-Brechot, V.; Bourhis, J. M.; Uversky, V. N.; Canard, B.; Longhi, S. *Proteins* **2006**, *62*, 24.
(147) Uversky, V. N. *Biochim. Biophys. Acta* **2013**, *1834*, 932.
(148) Alber, F.; Dokudovskaya, S.; Veenhoff, L. M.; Zhang, W.; Kipper, J.; Devos, D.; Suprapto, A.; Karni-Schmidt, O.; Williams, R.; Chait, B. T.; et al. *Nature* **2007**, *450*, 695.
(149) Yamada, J.; Phillips, J. L.; Patel, S.; Goldfien, G.; Calestagne-Morelli, A.; Huang, H.; Reza, R.; Acheson, J.; Krishnan, V. V.; Newsam, S.; et al. *Mol. Cell. Proteomics* **2010**, *9*, 2205.
(150) Denning, D. P.; Rexach, M. F. *Mol. Cell. Proteomics* **2007**, *6*, 272.
(151) Patel, S. S.; Belmont, B. J.; Sante, J. M.; Rexach, M. F. *Cell* **2007**, *129*, 83.
(152) Krishnan, V. V.; Lau, E. Y.; Yamada, J.; Denning, D. P.; Patel, S. S.; Colvin, M. E.; Rexach, M. F. *PLoS Comput. Biol.* **2008**, *4*, e1000145.
(153) Tompa, P. *Nat. Chem. Biol.* **2012**, *8*, 597.
(154) Yang, S.; Blachowicz, L.; Makowski, L.; Roux, B. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 15757.
(155) Wiesner, S.; Ogunjimi, A. A.; Wang, H. R.; Rotin, D.; Sicheri, F.; Wrana, J. L.; Forman-Kay, J. D. *Cell* **2007**, *130*, 651.
(156) Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. *FEBS Lett.* **2004**, *576*, 348.
(157) Lise, S.; Jones, D. T. *Proteins* **2005**, *58*, 144.
(158) Mao, A. H.; Lyle, N.; Pappu, R. V. *Biochem. J.* **2013**, *449*, 307.
(159) Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 16764.
(160) Mukhopadhyay, S.; Krishnan, R.; Lemke, E. A.; Lindquist, S.; Deniz, A. A. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 2649.
(161) Tran, H. T.; Mao, A.; Pappu, R. V. *J. Am. Chem. Soc.* **2008**, *130*, 7380.
(162) Teufel, D. P.; Johnson, C. M.; Lum, J. K.; Neuweiler, H. *J. Mol. Biol.* **2011**, *409*, 250.
(163) Papoian, G. A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14237.
(164) Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. *Arch. Biochem. Biophys.* **2008**, *469*, 132.
(165) Halfmann, R.; Alberti, S.; Krishnan, R.; Lyle, N.; O'Donnell, C. W.; King, O. D.; Berger, B.; Pappu, R. V.; Lindquist, S. *Mol. Cell* **2011**, *43*, 72.
(166) Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N.; et al. *Nucleic Acids Res.* **2007**, *35*, D786.
(167) Das, R. K.; Pappu, R. V. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 13392.
(168) Ferron, F.; Longhi, S.; Canard, B.; Karlin, D. *Proteins* **2006**, *65*, 1.
(169) Dosztanyi, Z.; Meszaros, B.; Simon, I. *Brief. Bioinform.* **2010**, *11*, 225.
(170) Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. *Proteins* **2003**, *52*, 573.
(171) Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. *Proteins* **2007**, *66*, 16.
(172) Laursen, B. S.; Kjaergaard, A. C.; Mortensen, K. K.; Hoffman, D. W.; Sperling-Petersen, H. U. *Protein Sci.* **2004**, *13*, 230.
(173) Edwards, Y. J.; Lobley, A. E.; Pentony, M. M.; Jones, D. T. *Genome Biol.* **2009**, *10*, R50.
(174) Galea, C. A.; High, A. A.; Obenauer, J. C.; Mishra, A.; Park, C. G.; Punta, M.; Schlessinger, A.; Ma, J.; Rost, B.; Slaughter, C. A.; et al. *J. Proteome Res.* **2009**, *8*, 211.
(175) Tompa, P.; Kalmar, L. *J. Mol. Biol.* **2010**, *403*, 346.
(176) Lobley, A.; Swindells, M. B.; Orengo, C. A.; Jones, D. T. *PLoS Comput. Biol.* **2007**, *3*, e162.
(177) Vuzman, D.; Azia, A.; Levy, Y. *J. Mol. Biol.* **2010**, *396*, 674.
(178) Magidovich, E.; Fleishman, S. J.; Yifrach, O. *Bioinformatics* **2006**, *22*, 1546.
(179) Jorda, J.; Xue, B.; Uversky, V. N.; Kajava, A. V. *FEBS J.* **2010**, *277*, 2673.
(180) Simon, M.; Hancock, J. M. *Genome Biol.* **2009**, *10*, R59.
(181) Matsushima, N.; Tanaka, T.; Kretsinger, R. H. *Protein Pept. Lett.* **2009**, *16*, 1297.

(182) Gerber, H. P.; Seipel, K.; Georgiev, O.; Hofferer, M.; Hug, M.; Rusconi, S.; Schaffner, W. *Science* **1994**, *263*, 808.
(183) Lobanov, M. Y.; Furletova, E. I.; Bogatyreva, N. S.; Roytberg, M. A.; Galzitskaya, O. V. *PLoS Comput. Biol.* **2010**, *6*, e1000958.
(184) Gsponer, J.; Babu, M. M. *Cell Rep.* **2012**, *2*, 1425.
(185) Michelitsch, M. D.; Weissman, J. S. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 11910.
(186) Fiumara, F.; Fioriti, L.; Kandel, E. R.; Hendrickson, W. A. *Cell* **2010**, *143*, 1121.
(187) Reijns, M. A.; Alexander, R. D.; Spiller, M. P.; Beggs, J. D. *J. Cell Sci.* **2008**, *121*, 2463.
(188) von Mikecz, A. *Trends Cell Biol.* **2009**, *19*, 685.
(189) DePace, A. H.; Santoso, A.; Hillner, P.; Weissman, J. S. *Cell* **1998**, *93*, 1241.
(190) Haynes, C.; Iakoucheva, L. M. *Nucleic Acids Res.* **2006**, *34*, 305.
(191) Shepard, P. J.; Hertel, K. J. *Genome Biol.* **2009**, *10*, 242.
(192) Ghosh, G.; Adams, J. A. *FEBS J.* **2011**, *278*, 587.
(193) Ponte, I.; Vila, R.; Suau, P. *Mol. Biol. Evol.* **2003**, *20*, 371.
(194) Wright, P. E.; Dyson, H. J.; Lerner, R. A. *Biochemistry* **1988**, *27*, 7167.
(195) Jahn, T. R.; Radford, S. E. *FEBS J.* **2005**, *272*, 5962.
(196) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789.
(197) Ma, B.; Nussinov, R. *Genome Biol.* **2009**, *10*, 204.
(198) Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. *Curr. Opin. Pharmacol.* **2010**, *10*, 715.
(199) Hammes, G. G.; Chang, Y. C.; Oas, T. G. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 13737.
(200) Song, J.; Guo, L. W.; Muradov, H.; Artemyev, N. O.; Ruoho, A. E.; Markley, J. L. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 1505.
(201) Eliezer, D.; Palmer, A. G., 3rd *Nature* **2007**, *447*, 920.
(202) Tompa, P.; Fuxreiter, M. *Trends Biochem. Sci.* **2008**, *33*, 2.
(203) Mittag, T.; Kay, L. E.; Forman-Kay, J. D. *J. Mol. Recognit.* **2010**, *23*, 105.
(204) Fuxreiter, M. *Mol. BioSyst.* **2012**, *8*, 168.
(205) Graham, T. A.; Ferkey, D. M.; Mao, F.; Kimelman, D.; Xu, W. *Nat. Struct. Biol.* **2001**, *8*, 1048.
(206) Renault, L.; Bugyi, B.; Carlier, M. F. *Trends Cell Biol.* **2008**, *18*, 494.
(207) Wang, Y.; Fisher, J. C.; Mathew, R.; Ou, L.; Otieno, S.; Sublet, J.; Xiao, L.; Chen, J.; Roussel, M. F.; Kriwacki, R. W. *Nat. Chem. Biol.* **2011**.
(208) Zor, T.; Mayr, B. M.; Dyson, H. J.; Montminy, M. R.; Wright, P. E. *J. Biol. Chem.* **2002**, *277*, 42241.
(209) Pometun, M. S.; Chekmenev, E. Y.; Wittebort, R. J. *J. Biol. Chem.* **2004**, *279*, 7982.
(210) Mittag, T.; Orlicky, S.; Choy, W. Y.; Tang, X.; Lin, H.; Sicheri, F.; Kay, L. E.; Tyers, M.; Forman-Kay, J. D. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17772.
(211) Fuxreiter, M.; Simon, I.; Bondos, S. *Trends Biochem. Sci.* **2011**, *36*, 415.
(212) Naud, J. F.; McDuff, F. O.; Sauve, S.; Montagne, M.; Webb, B. A.; Smith, S. P.; Chabot, B.; Lavigne, P. *Biochemistry* **2005**, *44*, 12746.
(213) Pufall, M. A.; Lee, G. M.; Nelson, M. L.; Kang, H. S.; Velyvis, A.; Kay, L. E.; McIntosh, L. P.; Graves, B. J. *Science* **2005**, *309*, 142.
(214) Stott, K.; Watson, M.; Howe, F. S.; Grossmann, J. G.; Thomas, J. O. *J. Mol. Biol.* **2010**, *403*, 706.
(215) Jonker, H. R.; Wechselberger, R. W.; Boelens, R.; Kaptein, R.; Folkers, G. E. *Biochemistry* **2006**, *45*, 5067.
(216) Uversky, V. N. *Chem. Soc. Rev.* **2011**, *40*, 1623.
(217) Goldman, N.; Thorne, J. L.; Jones, D. T. *Genetics* **1998**, *149*, 445.
(218) Bellay, J.; Michaut, M.; Kim, T.; Han, S.; Colak, R.; Myers, C. L.; Kim, P. M. *Mol. BioSyst.* **2012**, *8*, 185.
(219) Moesa, H. A.; Wakabayashi, S.; Nakai, K.; Patil, A. *Mol. BioSyst.* **2012**, *8*, 3262.
(220) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradovic, Z.; Dunker, A. K. *J. Mol. Biol.* **2002**, *323*, 573.
(221) Liu, J.; Perumal, N. B.; Oldfield, C. J.; Su, E. W.; Uversky, V. N.; Dunker, A. K. *Biochemistry* **2006**, *45*, 6873.
(222) Tantos, A.; Han, K. H.; Tompa, P. *Mol. Cell. Endocrinol.* **2012**, *348*, 457.
(223) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. *Curr. Opin. Struct. Biol.* **2011**, *21*, 432.
(224) Colak, R.; Kim, T.; Michaut, M.; Sun, M.; Irimia, M.; Bellay, J.; Myers, C. L.; Blencowe, B. J.; Kim, P. M. *PLoS Comput. Biol.* **2013**, *9*, e1003030.
(225) Tonikian, R.; Xin, X.; Toret, C. P.; Gfeller, D.; Landgraf, C.; Panni, S.; Paoluzi, S.; Castagnoli, L.; Currell, B.; Seshagiri, S.; et al. *PLoS Biol.* **2009**, *7*, e1000218.
(226) Dinkel, H.; Chica, C.; Via, A.; Gould, C. M.; Jensen, L. J.; Gibson, T. J.; Diella, F. *Nucleic Acids Res.* **2011**, *39*, D261.
(227) Beltrao, P.; Trinidad, J. C.; Fiedler, D.; Roguev, A.; Lim, W. A.; Shokat, K. M.; Burlingame, A. L.; Krogan, N. J. *PLoS Biol.* **2009**, *7*, e1000134.
(228) Ngo, J. C.; Giang, K.; Chakrabarti, S.; Ma, C. T.; Huynh, N.; Hagopian, J. C.; Dorrestein, P. C.; Fu, X. D.; Adams, J. A.; Ghosh, G. *Mol. Cell* **2008**, *29*, 563.
(229) Schad, E.; Tompa, P.; Hegyi, H. *Genome Biol.* **2011**, *12*, R120.

(230) Pancsa, R.; Tompa, P. *PLoS ONE* **2012**, *7*, e34687.
(231) Pavlovic-Lazetic, G. M.; Mitic, N. S.; Kovacevic, J. J.; Obradovic, Z.; Malkov, S. N.; Beljanski, M. V. *BMC Bioinformatics* **2011**, *12*, 66.
(232) Xue, B.; Williams, R. W.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N. *BMC Syst. Biol.* **2010**, *4 Suppl 1*, S1.
(233) Burra, P. V.; Kalmar, L.; Tompa, P. *PLoS ONE* **2010**, *5*, e12069.
(234) Tokuriki, N.; Oldfield, C. J.; Uversky, V. N.; Berezovsky, I. N.; Tawfik, D. S. *Trends Biochem. Sci.* **2009**, *34*, 53.
(235) Longhi, S. *Protein Pept. Lett.* **2010**, *17*, 930.
(236) Vidalain, P. O.; Tangy, F. *Microbes Infect.* **2010**, *12*, 1134.
(237) Xue, B.; Williams, R. W.; Oldfield, C. J.; Goh, G. K.; Dunker, A. K.; Uversky, V. N. *Protein Pept. Lett.* **2010**, *17*, 932.
(238) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, *37*, 215.
(239) Hegyi, H.; Buday, L.; Tompa, P. *PLoS Comput. Biol.* **2009**, *5*, e1000552.
(240) Uversky, V. N. *Front. Biosci.* **2009**, *14*, 5188.
(241) Uversky, V. N.; Oldfield, C. J.; Midic, U.; Xie, H.; Xue, B.; Vucetic, S.; Iakoucheva, L. M.; Obradovic, Z.; Dunker, A. K. *BMC Genomics* **2009**, *10 Suppl 1*, S7.
(242) Vavouri, T.; Semple, J. I.; Garcia-Verdugo, R.; Lehner, B. *Cell* **2009**, *138*, 198.
(243) Kriventseva, E. V.; Koch, I.; Apweiler, R.; Vingron, M.; Bork, P.; Gelfand, M. S.; Sunyaev, S. *Trends Genet.* **2003**, *19*, 124.
(244) Romero, P. R.; Zaidi, S.; Fang, Y. Y.; Uversky, V. N.; Radivojac, P.; Oldfield, C. J.; Cortese, M. S.; Sickmeier, M.; LeGall, T.; Obradovic, Z.; et al. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 8390.
(245) Hegyi, H.; Kalmar, L.; Horvath, T.; Tompa, P. *Nucleic Acids Res.* **2011**, *39*, 1208.
(246) Buljan, M.; Chalancon, G.; Eustermann, S.; Wagner, G. P.; Fuxreiter, M.; Bateman, A.; Babu, M. M. *Mol. Cell* **2012**, *46*, 871.
(247) Ellis, J. D.; Barrios-Rodiles, M.; Colak, R.; Irimia, M.; Kim, T.; Calarco, J. A.; Wang, X.; Pan, Q.; O'Hanlon, D.; Kim, P. M.; et al. *Mol. Cell* **2012**, *46*, 884.
(248) Buljan, M.; Chalancon, G.; Dunker, A. K.; Bateman, A.; Balaji, S.; Fuxreiter, M.; Babu, M. M. *Curr. Opin. Struct. Biol.* **2013**, *23*, 443.
(249) Reed, H. C.; Hoare, T.; Thomsen, S.; Weaver, T. A.; White, R. A.; Akam, M.; Alonso, C. R. *Genetics* **2010**, *184*, 745.
(250) Bondos, S. E.; Hsiao, H. C. *Adv. Exp. Med. Biol.* **2012**, *725*, 86.
(251) Merkin, J.; Russell, C.; Chen, P.; Burge, C. B. *Science* **2012**, *338*, 1593.
(252) Zarnack, K.; Konig, J.; Tajnik, M.; Martincorena, I.; Eustermann, S.; Stevant, I.; Reyes, A.; Anders, S.; Luscombe, N. M.; Ule, J. *Cell* **2013**, *152*, 453.
(253) Stein, A.; Aloy, P. *PLoS ONE* **2008**, *3*, e2524.
(254) Barbosa-Morais, N. L.; Irimia, M.; Pan, Q.; Xiong, H. Y.; Gueroussov, S.; Lee, L. J.; Slobodeniuc, V.; Kutter, C.; Watt, S.; Colak, R.; et al. *Science* **2012**, *338*, 1587.
(255) Weatheritt, R. J.; Davey, N. E.; Gibson, T. J. *Nucleic Acids Res.* **2012**, *40*, 7123.
(256) Liu, C. W.; Corboy, M. J.; DeMartino, G. N.; Thomas, P. J. *Science* **2003**, *299*, 408.
(257) Prakash, S.; Tian, L.; Ratliff, K. S.; Lehotzky, R. E.; Matouschek, A. *Nat. Struct. Mol. Biol.* **2004**, *11*, 830.
(258) Takeuchi, J.; Chen, H.; Coffino, P. *EMBO J.* **2007**, *26*, 123.
(259) Tompa, P.; Prilusky, J.; Silman, I.; Sussman, J. L. *Proteins* **2008**, *71*, 903.
(260) Schrader, E. K.; Harstad, K. G.; Matouschek, A. *Nat. Chem. Biol.* **2009**, *5*, 815.
(261) Tsvetkov, P.; Reuven, N.; Prives, C.; Shaul, Y. *J. Biol. Chem.* **2009**, *284*, 26234.
(262) Fishbain, S.; Prakash, S.; Herrig, A.; Elsasser, S.; Matouschek, A. *Nat. Commun.* **2011**, *2*, 192.
(263) Inobe, T.; Fishbain, S.; Prakash, S.; Matouschek, A. *Nat. Chem. Biol.* **2011**, *7*, 161.
(264) Ng, A. H.; Fang, N. N.; Comyn, S. A.; Gsponer, J.; Mayor, T. *Mol. Cell. Proteomics* **2013**.
(265) Ravid, T.; Hochstrasser, M. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 679.
(266) Deshaies, R. J.; Joazeiro, C. A. *Annu. Rev. Biochem.* **2009**, *78*, 399.
(267) Sharipo, A.; Imreh, M.; Leonchiks, A.; Imreh, S.; Masucci, M. G. *Nat. Med.* **1998**, *4*, 939.
(268) Zhang, M.; Coffino, P. *J. Biol. Chem.* **2004**, *279*, 8635.
(269) Tian, L.; Holmgren, R. A.; Matouschek, A. *Nat. Struct. Mol. Biol.* **2005**, *12*, 1045.
(270) Orlowski, M.; Wilk, S. *Arch. Biochem. Biophys.* **2003**, *415*, 1.
(271) Alvarez-Castelao, B.; Castano, J. G. *FEBS Lett.* **2005**, *579*, 4797.
(272) Asher, G.; Tsvetkov, P.; Kahana, C.; Shaul, Y. *Genes Dev.* **2005**, *19*, 316.
(273) Asher, G.; Reuven, N.; Shaul, Y. *BioEssays* **2006**, *28*, 844.
(274) Tsvetkov, P.; Asher, G.; Paz, A.; Reuven, N.; Sussman, J. L.; Silman, I.; Shaul, Y. *Proteins* **2008**, *70*, 1357.
(275) Wiggins, C. M.; Tsvetkov, P.; Johnson, M.; Joyce, C. L.; Lamb, C. A.; Bryant, N. J.; Komander, D.; Shaul, Y.; Cook, S. J. *J. Cell Sci.* **2011**, *124*, 969.
(276) Tsvetkov, P.; Reuven, N.; Shaul, Y. *Nat. Chem. Biol.* **2009**, *5*, 778.

(277) Kedersha, N.; Ivanov, P.; Anderson, P. *Trends Biochem. Sci.* **2013**.

(278) Kato, M.; Han, T. W.; Xie, S.; Shi, K.; Du, X.; Wu, L. C.; Mirzaei, H.; Goldsmith, E. J.; Longgood, J.; Pei, J.; et al. *Cell* **2012**, *149*, 753.

(279) Tompa, P. *Intrinsically Disordered Proteins* **2013**, *1*, e24068.

(280) Guettler, S.; LaRose, J.; Petsalaki, E.; Gish, G.; Scotter, A.; Pawson, T.; Rottapel, R.; Sicheri, F. *Cell* **2011**, *147*, 1340.

(281) Jiang, K.; Toedt, G.; Montenegro Gouveia, S.; Davey, N. E.; Hua, S.; van der Vaart, B.; Grigoriev, I.; Larsen, J.; Pedersen, L. B.; Bezstarosti, K.; et al. *Curr. Biol.* **2012**, *22*, 1800.

(282) Obenauer, J. C.; Cantley, L. C.; Yaffe, M. B. *Nucleic Acids Res.* **2003**, *31*, 3635.

(283) Via, A.; Gould, C. M.; Gemund, C.; Gibson, T. J.; Helmer-Citterich, M. *BMC Bioinformatics* **2009**, *10*, 351.

(284) Schiller, M. R. *Curr. Protoc. Protein Sci.* **2007**, *Chapter 2*, Unit 2 12.

(285) Chica, C.; Labarga, A.; Gould, C. M.; Lopez, R.; Gibson, T. J. *BMC Bioinformatics* **2008**, *9*, 229.

(286) Gould, C. M.; Diella, F.; Via, A.; Puntervoll, P.; Gemund, C.; Chabanis-Davidson, S.; Michael, S.; Sayadi, A.; Bryne, J. C.; Chica, C.; et al. *Nucleic Acids Res.* **2010**, *38*, D167.

(287) Linding, R.; Jensen, L. J.; Ostheimer, G. J.; van Vugt, M. A.; Jorgensen, C.; Miron, I. M.; Diella, F.; Colwill, K.; Taylor, L.; Elder, K.; et al. *Cell* **2007**, *129*, 1415.

(288) Rajasekaran, S.; Merlin, J. C.; Kundeti, V.; Mi, T.; Oommen, A.; Vyas, J.; Alaniz, I.; Chung, K.; Chowdhury, F.; Deverasatty, S.; et al. *Proteins* **2011**, *79*, 153.

(289) Davey, N. E.; Cowan, J. L.; Shields, D. C.; Gibson, T. J.; Coldwell, M. J.; Edwards, R. J. *Nucleic Acids Res.* **2012**, *40*, 10628.

(290) Neduva, V.; Linding, R.; Su-Angrand, I.; Stark, A.; de Masi, F.; Gibson, T. J.; Lewis, J.; Serrano, L.; Russell, R. B. *PLoS Biol.* **2005**, *3*, e405.

(291) Davey, N. E.; Haslam, N. J.; Shields, D. C.; Edwards, R. J. *Nucleic Acids Res.* **2010**, *38*, W534.

(292) Disfani, F. M.; Hsu, W. L.; Mizianty, M. J.; Oldfield, C. J.; Xue, B.; Dunker, A. K.; Uversky, V. N.; Kurgan, L. *Bioinformatics* **2012**, *28*, i75.

(293) Dosztanyi, Z.; Meszaros, B.; Simon, I. *Bioinformatics* **2009**, *25*, 2745.

(294) Meszaros, B.; Simon, I.; Dosztanyi, Z. *PLoS Comput. Biol.* **2009**, *5*, e1000376.

(295) Fukuchi, S.; Sakamoto, S.; Nobe, Y.; Murakami, S. D.; Amemiya, T.; Hosoda, K.; Koike, R.; Hiroaki, H.; Ota, M. *Nucleic Acids Res.* **2012**, *40*, D507.

(296) Lobley, A. E.; Nugent, T.; Orengo, C. A.; Jones, D. T. *Nucleic Acids Res.* **2008**, *36*, W297.

(297) Cozzetto, D.; Jones, D. T. *Curr. Opin. Struct. Biol.* **2013**, *23*, 467.

(298) Minneci, F.; Piovesan, D.; Cozzetto, D.; Jones, D. T. *PLoS ONE* **2013**, *8*, e63754.

(299) Neduva, V.; Russell, R. B. *Curr. Opin. Biotechnol.* **2006**, *17*, 465.

(300) Vitalis, A.; Pappu, R. V. *J. Comput. Chem.* **2009**, *30*, 673.

(301) Fisher, C. K.; Stultz, C. M. *J. Am. Chem. Soc.* **2011**, *133*, 10022.

(302) Lyle, N.; K. Das, R.; Pappu, R. V. *J. Chem. Phys.* **2013**, *139*, 121907.

(303) Pawson, T. *Curr. Opin. Cell Biol.* **2007**, *19*, 112.

(304) Vucetic, S.; Obradovic, Z.; Vacic, V.; Radivojac, P.; Peng, K.; Iakoucheva, L. M.; Cortese, M. S.; Lawson, J. D.; Brown, C. J.; Sikes, J. G.; et al. *Bioinformatics* **2005**, *21*, 137.

(305) Di Domenico, T.; Walsh, I.; Martin, A. J.; Tosatto, S. C. *Bioinformatics* **2012**, *28*, 2080.

(306) Hornbeck, P. V.; Kornhauser, J. M.; Tkachev, S.; Zhang, B.; Skrzypek, E.; Murray, B.; Latham, V.; Sullivan, M. *Nucleic Acids Res.* **2012**, *40*, D261.

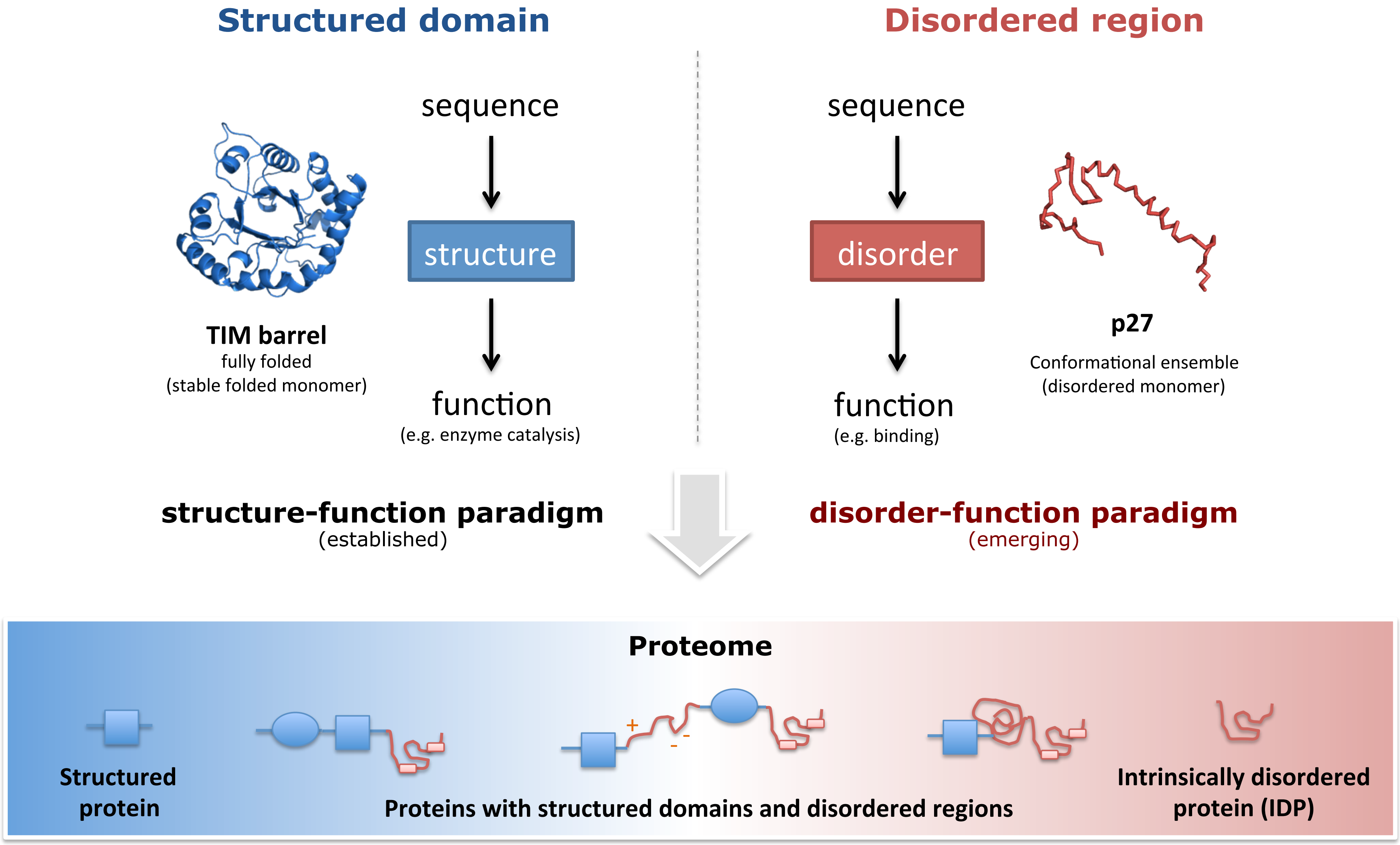
(307) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607.

(308) Deng, X.; Eickholt, J.; Cheng, J. *BMC Bioinformatics* **2009**, *10*, 436.

(309) Mayrose, I.; Graur, D.; Ben-Tal, N.; Pupko, T. *Mol. Biol. Evol.* **2004**, *21*, 1781.

(310) Chen, S. C.; Chuang, T. J.; Li, W. H. *Mol. Biol. Evol.* **2011**.

(311) Kajava, A. V.; Baxa, U.; Steven, A. C. *FASEB J.* **2010**, *24*, 1311.



Structured domain

Disordered region

sequence → structure → function

(e.g. enzyme catalysis)

TIM barrel

fully folded

(stable folded monomer)

sequence → disorder → function

(e.g. binding)

p27

Conformational ensemble

(disordered monomer)

structure-function paradigm

(established)

disorder-function paradigm

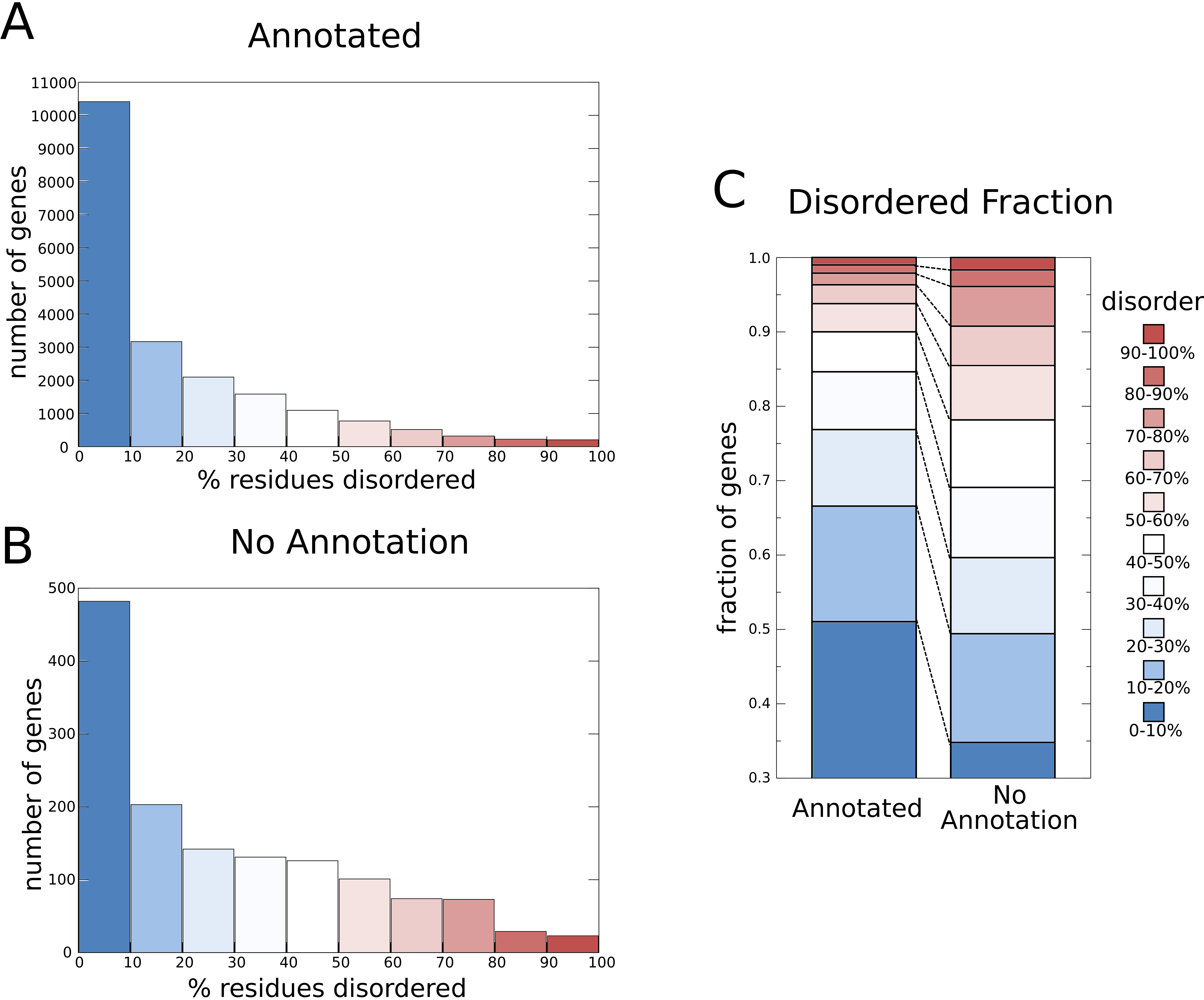
(emerging)

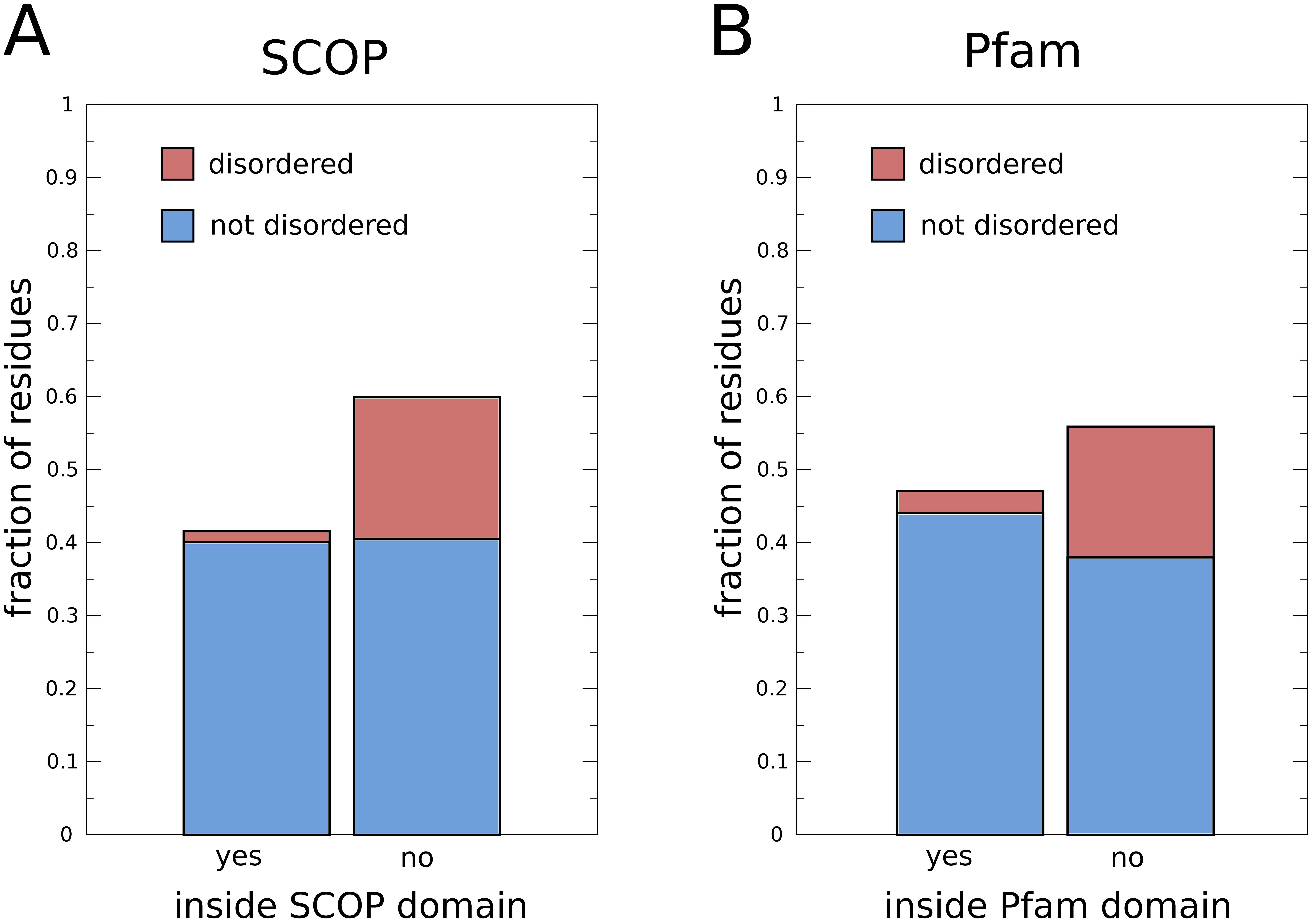
Proteome

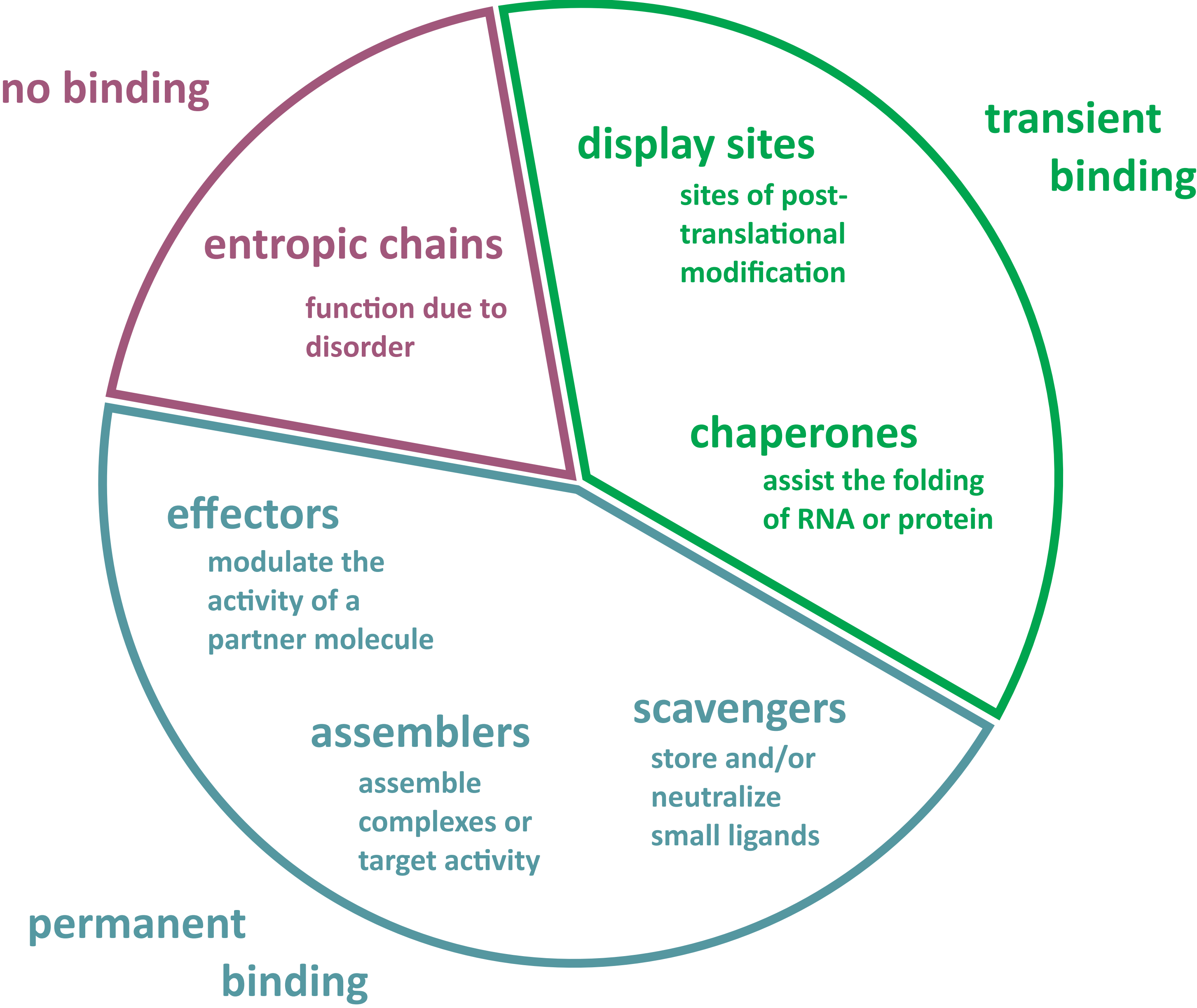
Structured protein

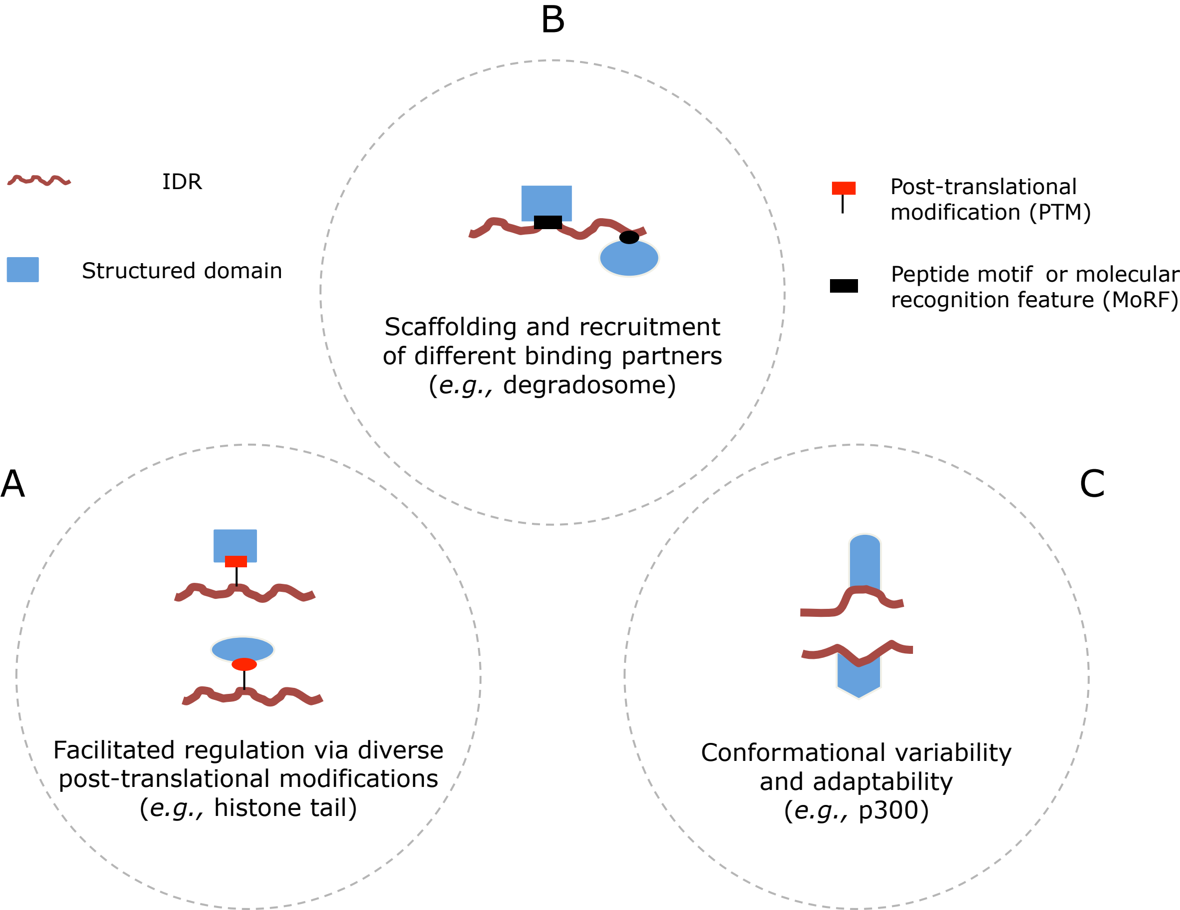
Proteins with structured domains and disordered regions

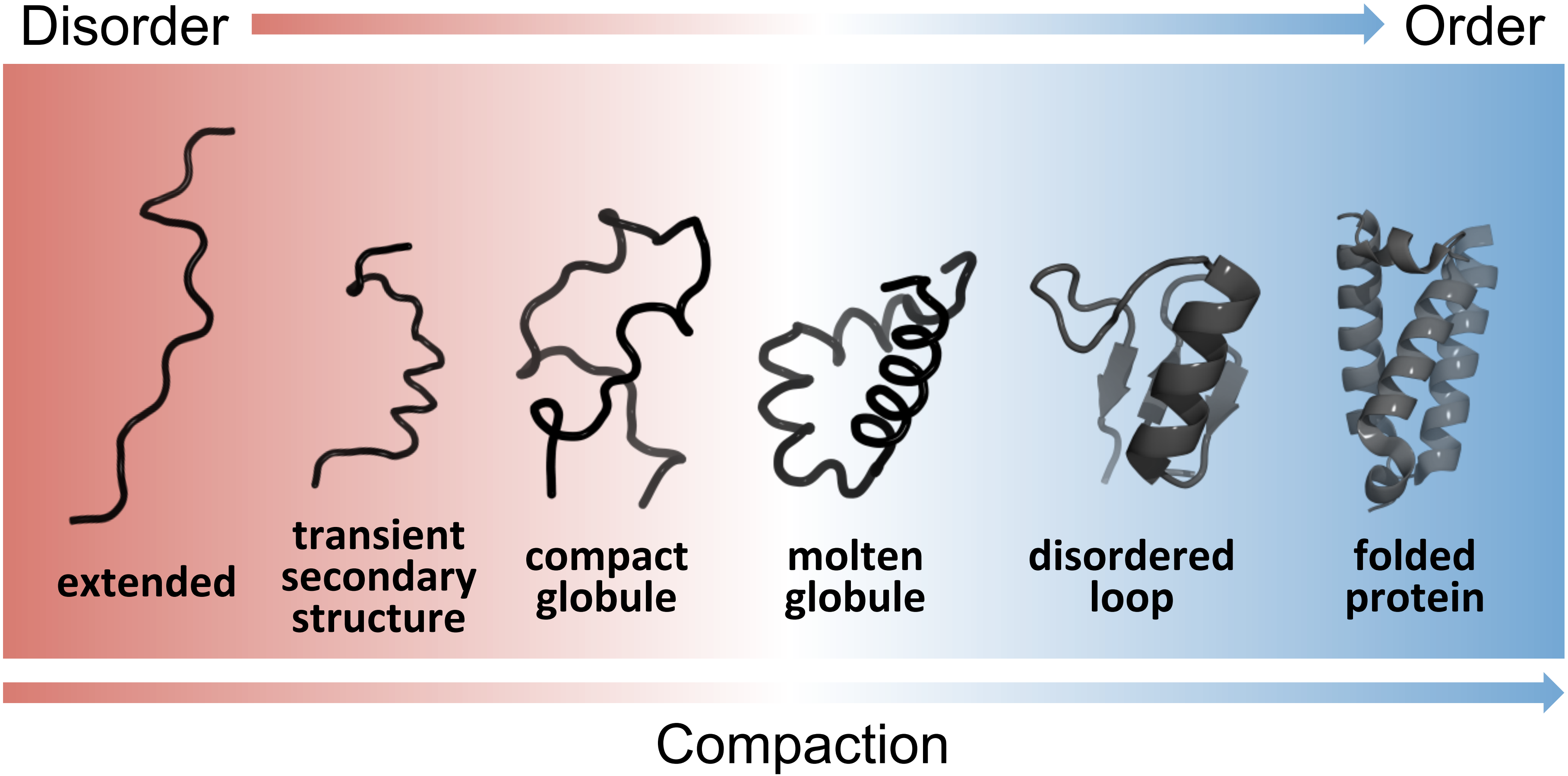
Intrinsically disordered protein (IDP)

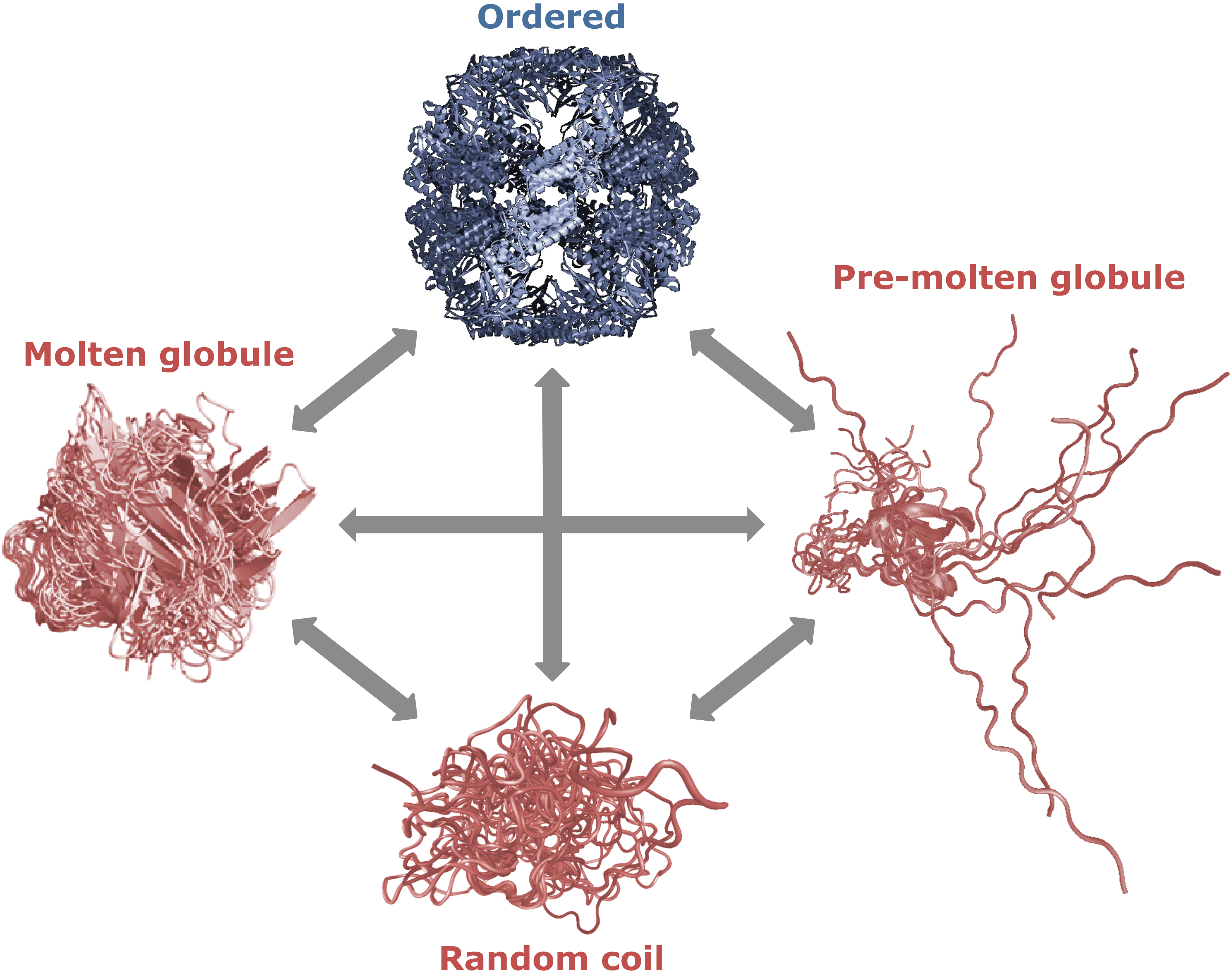

A
Annotated
number of genes
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70
80
90
100
% residues disordered
B
No Annotation
number of genes
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
90
100
% residues disordered
C
Disordered Fraction
fraction of genes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Annotated
No Annotation
disorder
90-100%
80-90%
70-80%
60-70%
50-60%
40-50%
30-40%
20-30%
10-20%
0-10%


A
SCOP
disordered
not disordered
fraction of residues
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
yes
no
inside SCOP domain
B
Pfam
disordered
not disordered
fraction of residues
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
yes
no
inside Pfam domain

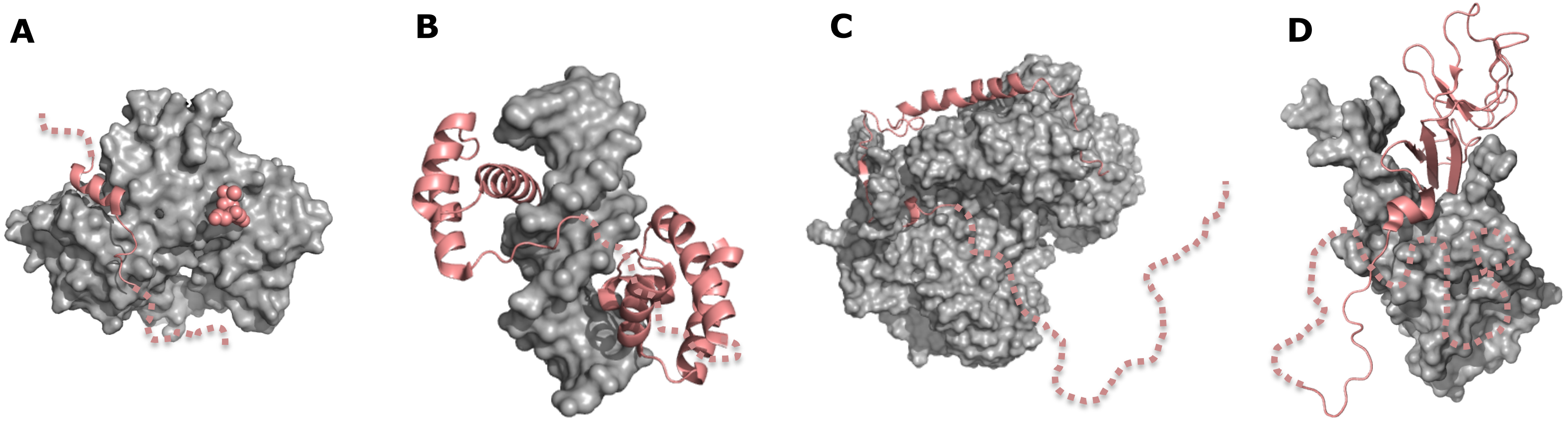

no binding
entropic chains
function due to disorder
transient binding
display sites
sites of post-translational modification
chaperones
assist the folding of RNA or protein
effectors
modulate the activity of a partner molecule
assemblers
assemble complexes or target activity
scavengers
store and/or neutralize small ligands
permanent binding


B
IDR
Structured domain
Post-translational modification (PTM)
Peptide motif or molecular recognition feature (MoRF)
Scaffolding and recruitment of different binding partners (*e.g.*, degradosome)
A
C
Facilitated regulation via diverse post-translational modifications (*e.g.*, histone tail)
Conformational variability and adaptability (*e.g.*, p300)

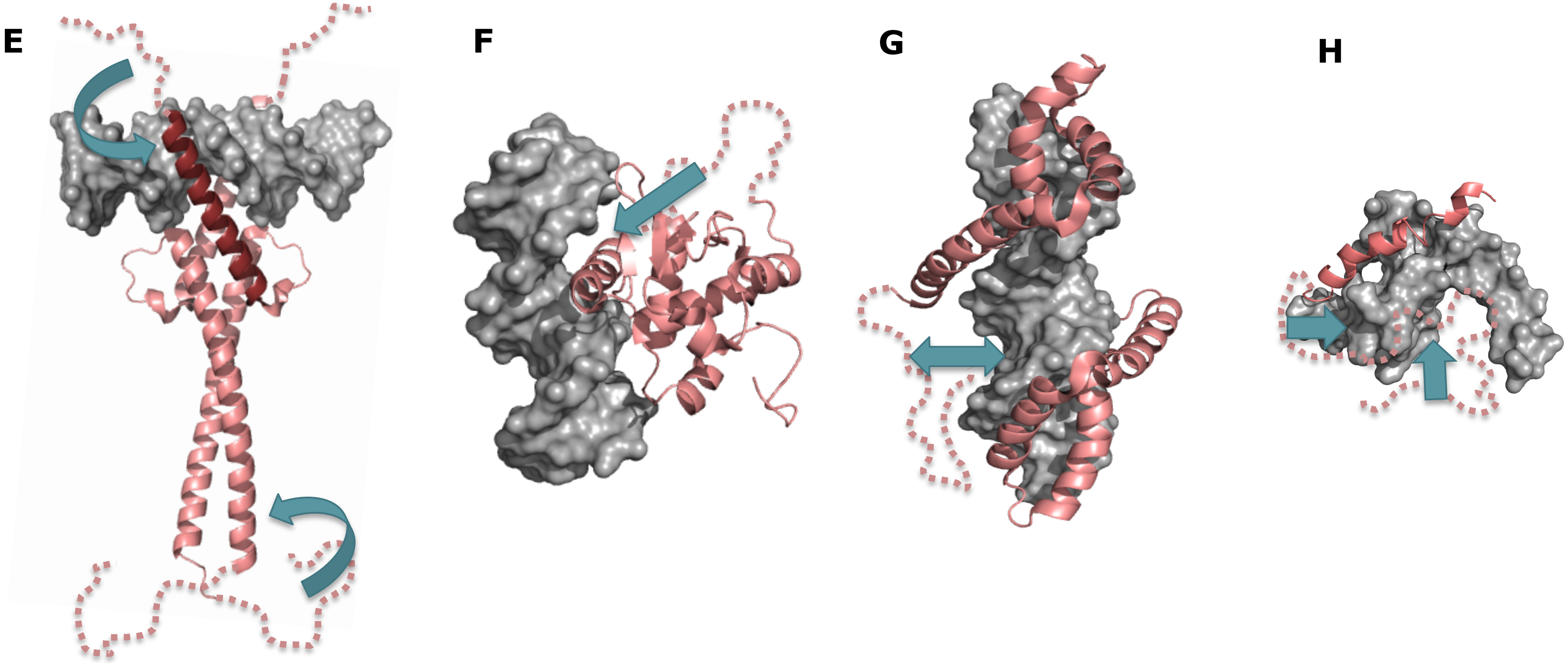

Disorder
Order
extended
transient secondary structure
compact globule
molten globule
disordered loop
folded protein
Compaction

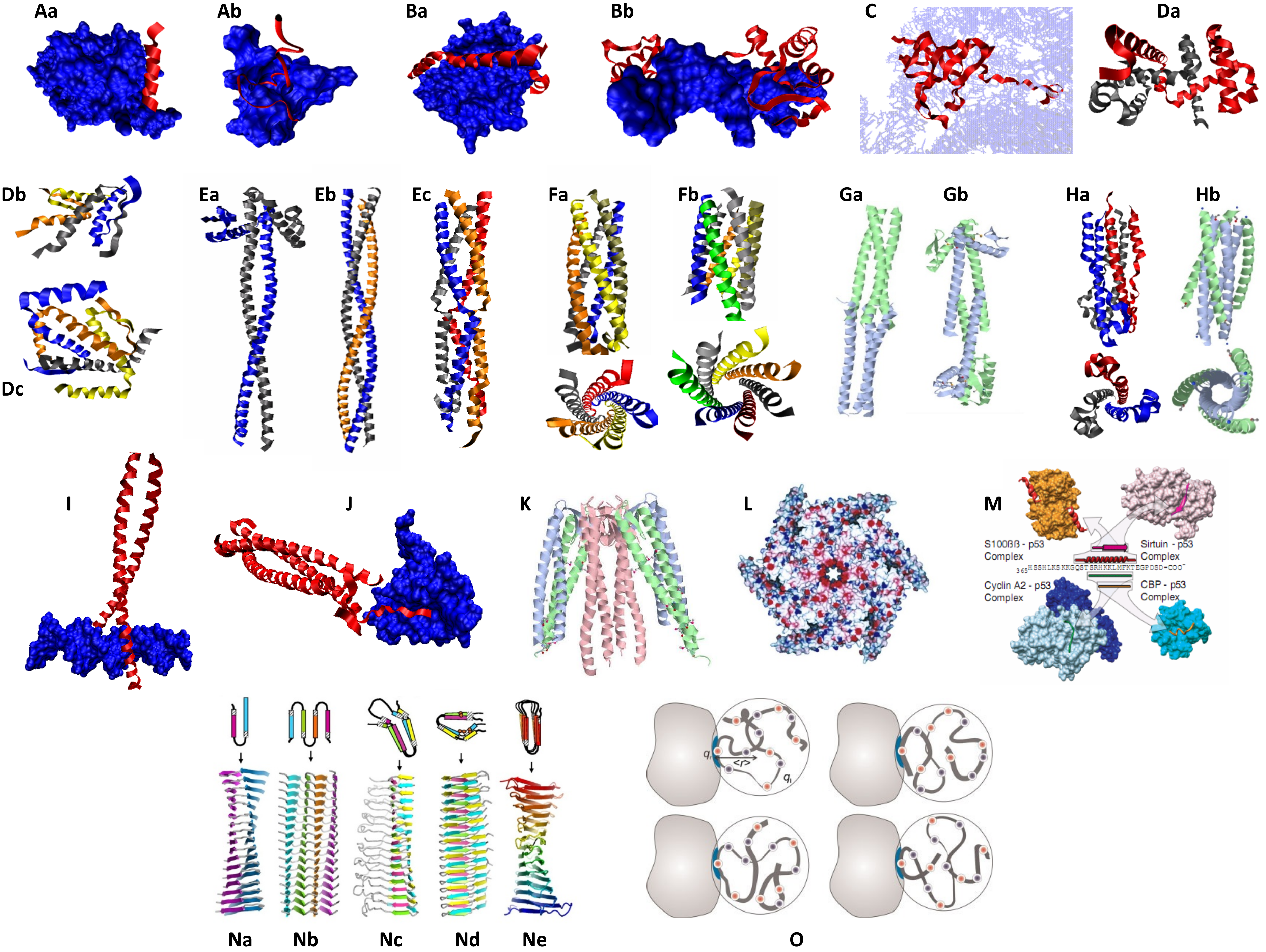

Ordered
Molten globule
Pre-molten globule
Random coil

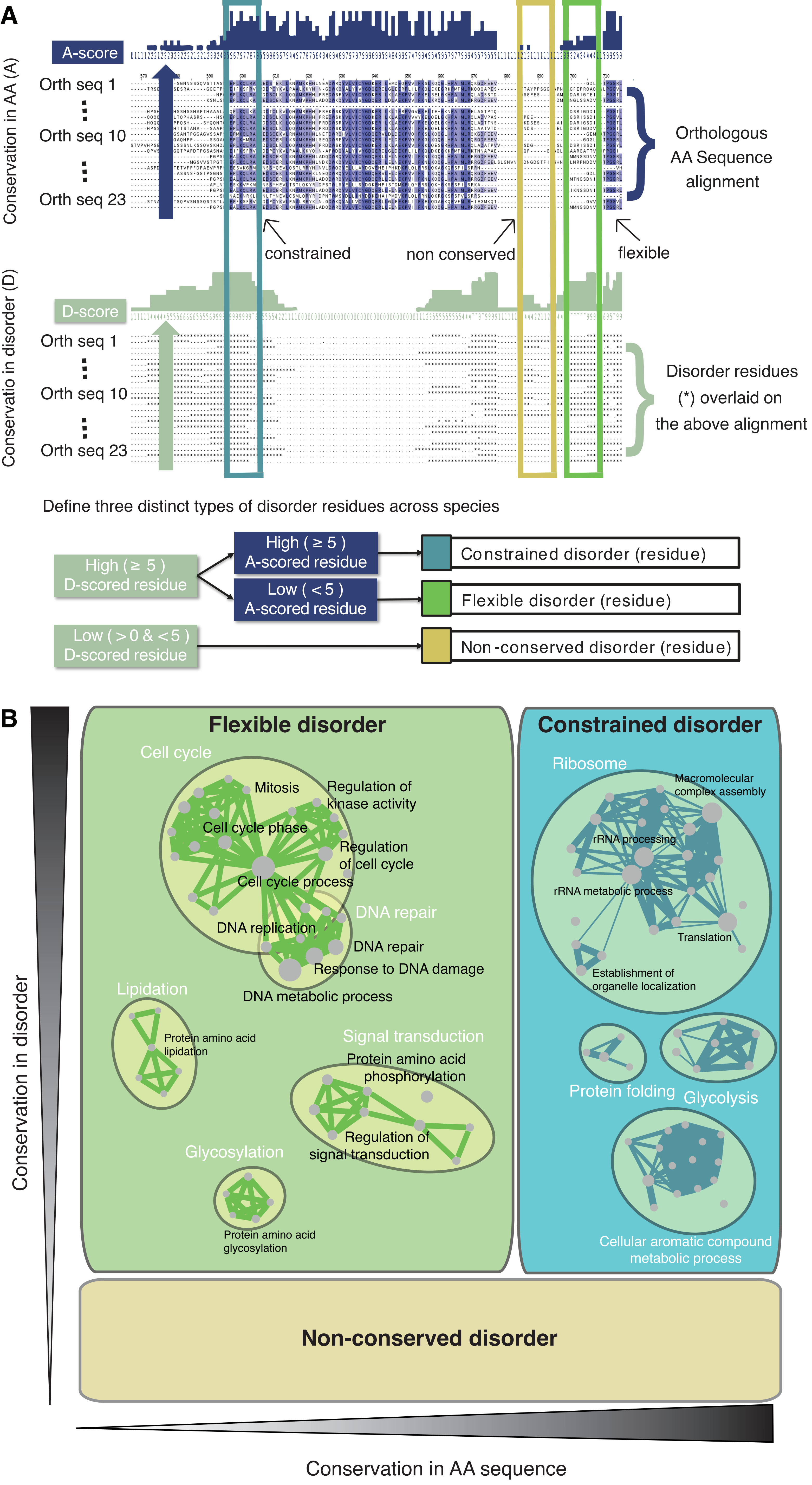
Topological categories of fuzzy complexes

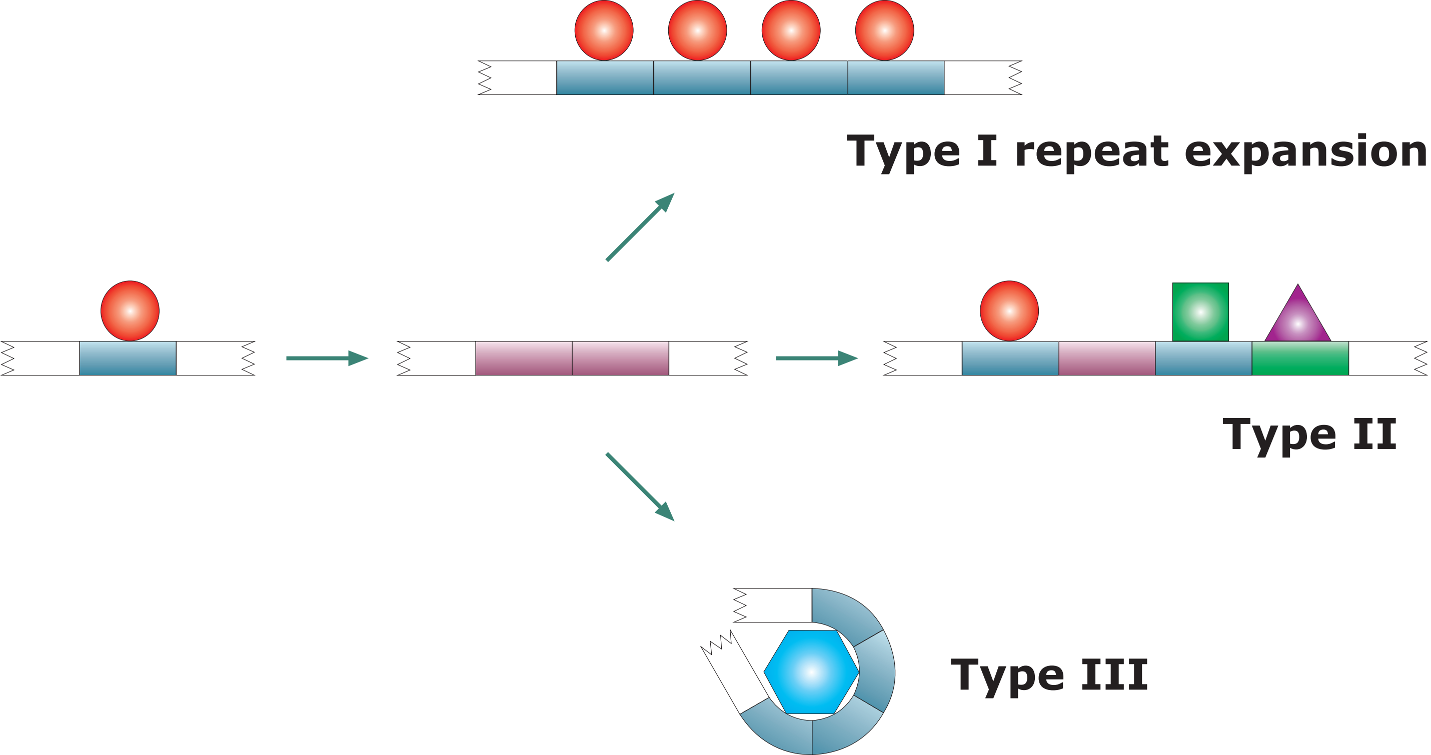


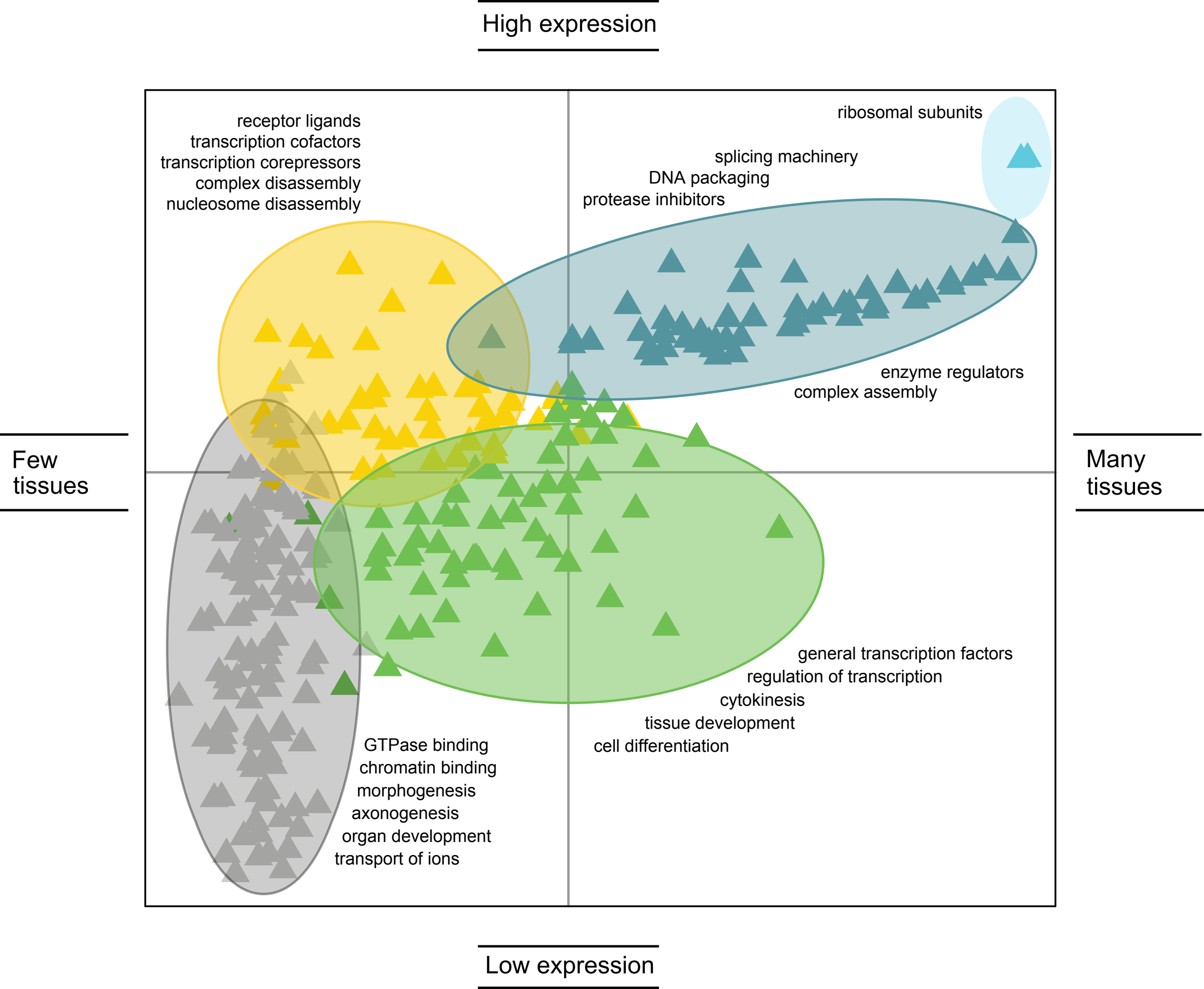
Categories of fuzzy complexes by mechanisms

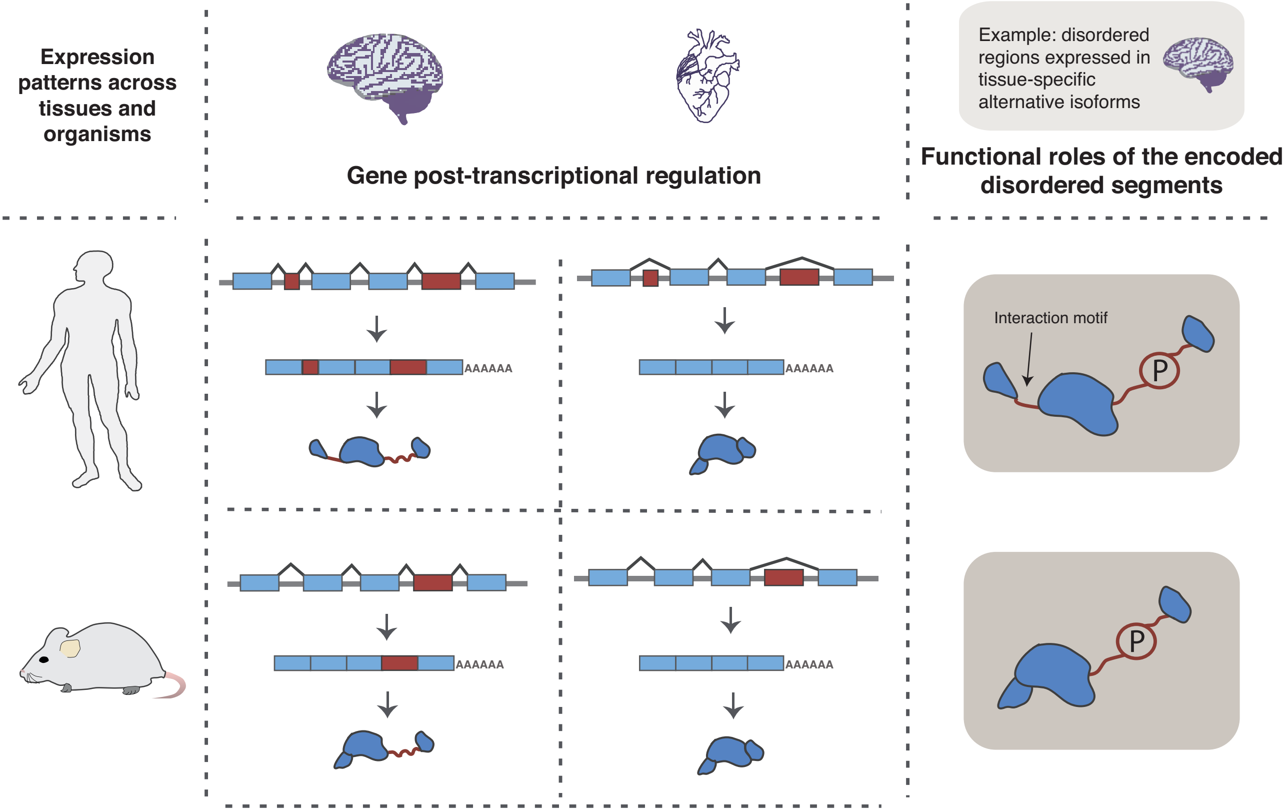



Aa
Ab
Ba
Bb
C
Da
Db
Dc
Ea
Eb
Ec
Fa
Fb
Ga
Gb
Ha
Hb
I
J
K
L
M
S100ßß - p53
Complex
Sirtuin - p53
Complex
365 HSSHLKSKKGQSTSRHKKLMFKTEGPDSD-COO⁻
Cyclin A2 - p53
Complex
CBP - p53
Complex
Na
Nb
Nc
Nd
Ne
q
<r>
q
O


A
A-score
Conservation in AA (A)
Orth seq 1
Orth seq 10
Orth seq 23
Orthologous
AA Sequence
alignment
constrained
non conserved
flexible
D-score
Conservatio in disorder (D)
Orth seq 1
Orth seq 10
Orth seq 23
Disorder residues
(*) overlaid on
the above alignment
Define three distinct types of disorder residues across species
High (≥ 5)
D-scored residue
High (≥ 5)
A-scored residue
Low (< 5)
A-scored residue
Low (> 0 & < 5)
D-scored residue
Constrained disorder (residue)
Flexible disorder (residue)
Non-conserved disorder (residue)
B
Conservation in disorder
Flexible disorder
Cell cycle
Mitosis
Regulation of
kinase activity
Cell cycle phase
Regulation
of cell cycle
Cell cycle process
DNA repair
DNA replication
DNA repair
Response to DNA damage
DNA metabolic process
Lipidation
Protein amino acid
lipidation
Signal transduction
Protein amino acid
phosphorylation
Regulation of
signal transduction
Glycosylation
Protein amino acid
glycosylation
Constrained disorder
Ribosome
Macromolecular
complex assembly
rRNA processing
rRNA metabolic process
Translation
Establishment of
organelle localization
Protein folding
Glycolysis
Cellular aromatic compound
metabolic process
Non-conserved disorder
Conservation in AA sequence


Type I repeat expansion
Type II
Type III


High expression
Few tissues
Many tissues
Low expression
receptor ligands
transcription cofactors
transcription corepressors
complex disassembly
nucleosome disassembly
ribosomal subunits
splicing machinery
DNA packaging
protease inhibitors
enzyme regulators
complex assembly
general transcription factors
regulation of transcription
cytokinesis
tissue development
cell differentiation
GTPase binding
chromatin binding
morphogenesis
axonogenesis
organ development
transport of ions

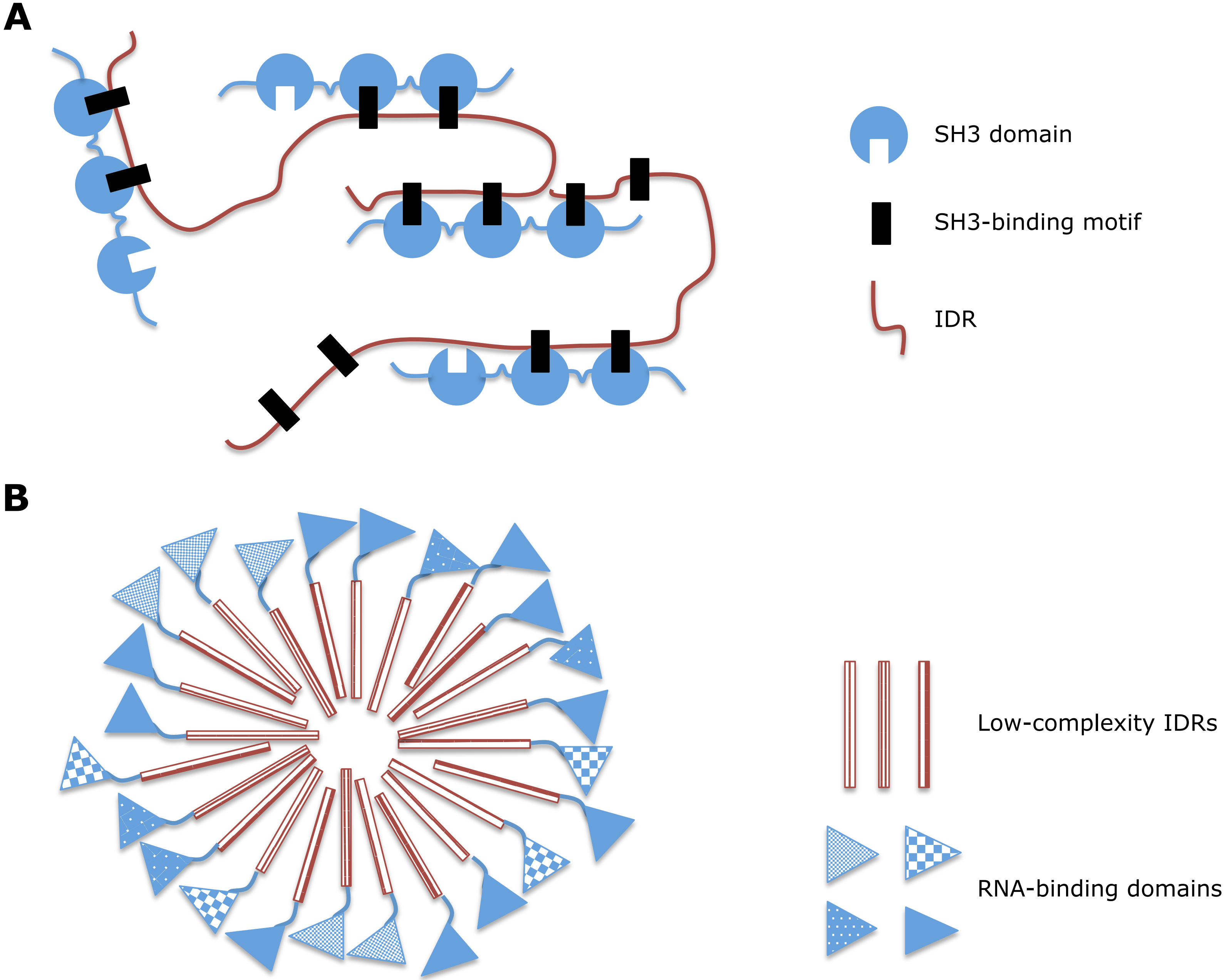


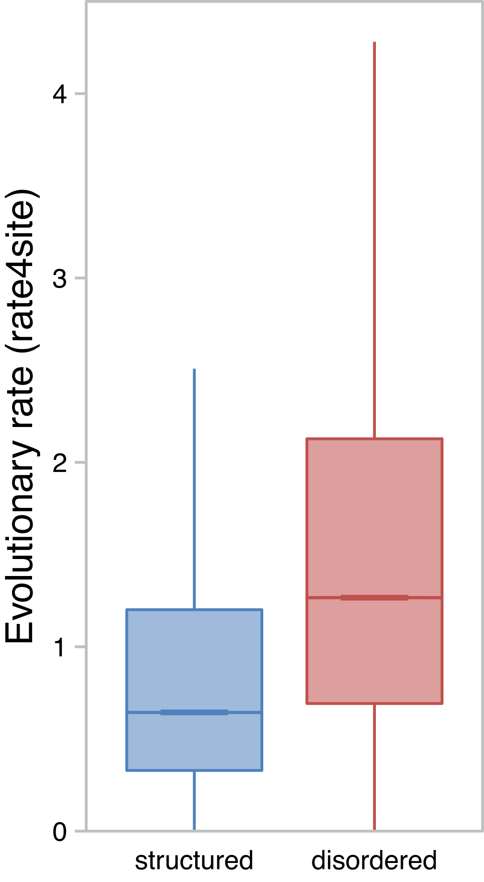
Expression patterns across tissues and organisms

Gene post-transcriptional regulation

Example: disordered regions expressed in tissue-specific alternative isoforms

Functional roles of the encoded disordered segments


A
SH3 domain
SH3-binding motif
IDR
B
Low-complexity IDRs
RNA-binding domains


Evolutionary rate (rate4site)
0
1
2
3
4
structured
disordered