CHAPTER

Introduction to intrinsically disordered proteins and regions 1

Christopher J. Oldfield*, Vladimir N. Uversky[†,‡,1], A. Keith Dunker[§,1], Lukasz Kurgan*,[1]

Department of Computer Science, Virginia Commonwealth University, Richmond, VA, United States Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, United States[†] Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow, Russia[‡] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, United States[§]*

[1]Corresponding authors. E-mails: vuversky@health.usf.edu (Vladimir N. Uversky), kedunker@iu.edu (A. Keith Dunker), and lkurgan@vcu.edu (Lukasz Kurgan)

1.1 Introduction

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) lack a stable structure in isolation, and instead exist as an ensemble of conformations that varies over time and populations *(1–3)*. The term "intrinsically disordered" was originally borrowed from the term "disorder" as used by X-ray crystallographers. This term refers to portions of a structure that are not resolved due to variable or fluctuating positions with the crystal lattice. "Intrinsic" is meant to reflect that structure and disordered states are encoded in a protein sequence, so these structural properties are intrinsic to the sequence.

A given protein may have one, many, or no IDRs, and in some cases entire proteins are intrinsically disordered. For example, the X-ray crystal structure of the protein Bcl-xL has several intrinsically disordered regions (Fig. 1.1). Like many proteins, Bcl-xL has an IDR at one of its termini, the C- terminus. In general, terminal IDRs can vary greatly in length, from a few to hundreds of residues, and can be present at both termini. Bcl-xL also has a large IDR that occurs within the bounds of its structured domains. These domain-internal disordered regions may also vary greatly in length. Though they are typically shorter than terminal disordered regions, there seems to be no physical limit to their length *(5)*. The other possible sequence location of IDRs is between domains, so-called linker regions. Linkers can be any length, from just a few residues to hundreds to a thousand residues, as in the case of BRCA1, which has a 1500 residue disordered region linking ordered domains at its N- and C- termini *(6)*.

Intrinsically Disordered Proteins. https://doi.org/10.1016/B978-0-12-816348-1.00001-6

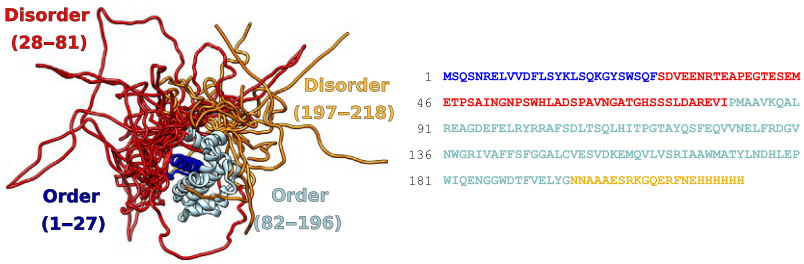

FIG. 1.1

Representation of disordered regions in the context of a protein's structure. The X-ray crystal structure of Bcl-xL *(4)* (*light* and *dark blue* ribbons) is shown along with several possible conformations of its two IDRs (*red* and *orange* worms). IDR conformations are randomly generated to scale. The sequence of Bcl-xL is also shown, with colors matching the structure diagram.

IDRs enable a wide variety of biological processes via mechanisms that are distinct from their structured counterparts. Broad analysis relating protein function to intrinsic disorder reveals a central role for IDRs in molecular recognition and cellular regulation *(7)*. Often, structured and disordered regions work in concert, as in the case of the lac repressor *(8)*. Intrinsic disorder enables complex regulation using concerted molecular recognition, posttranslational modification, and alternative splicing. The Nfat family of proteins exemplifies this phenomenon *(9)*. Intrinsic disorder also plays an important role in the inhibition and disinhibition of central cellular regulation, which is exemplified by p27 and related proteins *(10–13)*. The diverse biophysical properties of IDRs enable their myriad functions. While all IDRs are similar in that they lack an intrinsic stable structure, the biophysical properties of IDRs vary to all extremes. The biophysical and functional properties of IDRs are reflected in their composition, sequence complexity, and conservation. These sequence properties have been leveraged to design algorithms that predict intrinsic disorder from a protein sequence with high accuracy.

1.2 Biological processes associated with IDPs and IDRs

For structured proteins (proteins that do not include IDRs), the widely recognized concept that drives functional characterization is that sequence encodes structure and structure encodes function. The widespread use of disorder for biological function—revealed at least in part by computational studies—has led to a coequal summarizing concept, namely sequence encodes an IDP ensemble, the IDP ensemble encodes function.

Two approaches have been commonly used for correlating biological processes with proteins that may be structured, that may be IDPs, or that may utilize IDRs. The first approach is focused experimentation on particular proteins to determine their roles in cellular processes *(14)* and in structure-function relationships *(15–17)*.

Structure-function relationships can reveal that a particular function involves IDPs or IDRs or even the interactions between IDPs and structured protein partners. Here, we describe a few important proteins and their associated biological processes, highlighting results that suggest intrinsic disorder plays important roles in these processes. The second approach is broad computational studies that discover novel associations between the function and structure of IDRs, followed by further work to verify the implications of these associations *(7, 18–21)*. Because many if not most proteins contain both structured and disordered regions, often it is not clear whether the key molecular function underlying the indicated biological process actually utilizes the IDR or the structured region or even the collaboration of their structured and intrinsically disordered regions. Thus, such bioinformatics approaches need to be followed up with further studies to remove this ambiguity.

Identifying structure-function relationships for IDPs and proteins with IDRs is a major challenge. Two similar bioinformatics reports used computational approaches to estimate structure-function relationships on a large scale of thousands of proteins. The first report used the data on the model organism, yeast, *(18)* and the second used the Swiss-Prot (now Uniprot) database *(7)*. Their respective results are given in Table 1.1 (left column, *(18)*, right column, *(7)*). Both studies used very similar statistical evaluations to rank the annotated biological processes in the respective databases with regard to whether the functions were carried out by structured proteins or IDPs based on predictions, and the results of both studies were very similar overall. The reader should consult the two papers to understand the similarities and differences in the methods used. The focus here is on the similarities and differences in the reported results as given in Table 1.1.

First, consider the biological processes associated with the structured proteins. The biological processes associated with energy pathways (# 1, left column) for the yeast proteins correlate with the SwissProt protein processes related to electron transport (#4, right column), aromatic hydrocarbon catabolism (#6 right column), glycolysis (#7, right column), and carbohydrate metabolism (#10, right column). The yeast proteins have the annotation "process unknown" (#2, left column). Such a result is not surprising because even highly studied organisms such as S. cerevisiae and E. coli have many proteins of unknown function. On the other hand, all the proteins in SwissProt have been validated in the laboratory, so proteins of unknown function are rare in this database. The biosynthesis process (#3, left column) for the yeast proteins correlates with the SwissProt protein processes GMP biosynthesis (#1, right column), amino acid biosynthesis (#2, right column), lipid A biosynthesis (#5, right column), purine biosynthesis (#8, right column), and pyrimidine biosynthesis (#9, right column). These data show that the biological processes tend to be lumped in the first study (left column) and analyzed at a finer resolution level in the second study (right column). The reason for this is that the yeast proteome has only about 5300 to 5400 proteins *(26)*, whereas, at the time of the second study, the SwissProt database already contained more than 200,000 proteins *(7)*. The larger amount of data allowed a finer-grain evaluation of the biological processes associated with structure and disorder in the second study. Because the first study focused on IDPs, the authors chose to only present three biological processes associated with low predictions of disorder

Table 1.1 Biological processes predicted to utilize IDPs or structured proteins (SPs)

Processes associated with IDPs[a]	Processes associated with IDPs[b]
1. Ty transposition[c]	1. Differentiation
2. Development[d]	2. Transcription
3. Morphogenesis	3. Transcription regulation
4. Protein phosphorylation	4. Spermatogenesis
5. Regulation of transcription	5. DNA condensation
6. Transcription	6. Cell cycle
7. DNA packaging	7. mRNA processing
8. Signal transduction	8. mRNA splicing
9. Actin cytoskeleton biogenesis	9. Mitosis
10. Pseudohyphal growth[e]	10. Apoptosis
Processes associated with SPs[a]	**Processes associated with SPs[b]**
1. Energy pathways	1. GMP biosynthesis
2. Process unknown	2. Amino acid biosynthesis
3. Biosynthesis	3. Transport
4. Not given	4. Electron transport
5. Not given	5. Lipid A biosynthesis
6. Not given	6. Aromatic hydrocarbon catabolism
7. Not given	7. Glycolysis
8. Not given	8. Purine biosynthesis
9. Not given	9. Pyrimidine biosynthesis
10. Not given	10. Carbohydrate metabolism

[a] *These results are from Ward et al. (18).*
[b] *These results are from Xie et al. (7).*
[c] *Ty Transposition: associated with a Ty Transposable Element that resembles a primitive retrovirus, specific to yeast (22).*
[d] *Development: cell polarity (23), budding (24), and pseudohyphal growth (25) are among the developmental processes studied in yeast.*
[e] *Pseudohyphal growth: A pattern of cell growth that occurs in conditions of nitrogen limitation and an abundant fermentable carbon source. Cells become elongated, switch to a unipolar budding pattern, remain physically attached to each other, and invade the growth substrate (copied from the Gene Ontology Database).*

(e.g., with structured proteins), and only two of these processes provide useful annotations. Nevertheless, these two biological processes in the first study correlate with 9 of the 10 biological processes in the second study due to the respective lumping and splitting of the two studies. An interesting feature of both studies is that the proteins involved in the biological processes associated with structured proteins are all either membrane proteins or enzymes, and both of these categories are known to be structured.

With respect to the biological processes associated with disorder, the first study identified two IDP-associated processes specific to yeast, that is, Ty transposition

(#1, left) and pseudohydral growth (#10, left). Likewise, the second study identified three IDP processes specifically associated with multicellularity, that is, differentiation (#1, right), spermatogenesis (#4, right), and apoptosis (#10, right). The first study also identified five processes not observed among the top 10 in the second study, that is, development (#2, left), morphogenesis (#3, left), protein phosphorylation (#4, left); signal transduction (#8, left), and actin cytoskeleton biogenesis (#9, left), whereas the second study identified four processes not observed among the top 10 in the first, that is cell cycle (#6, right), mRNA processing (#7, right), mRNA splicing (#8,right), and mitosis (#9, right). The development and morphogenesis identified in the first study are features that are found to be associated with differentiation identified in the second, so it is likely that these unmatched categories likely utilize many homologous proteins. The cell cycle identified in the second study but not in the top 10 of the first is identified (at #20) in the first study. Protein phosphorylation and signal transduction, which are identified in the first study but not in the second, are both known to be common features of IDPs, so the reason for the lack of a match for these examples is unclear. The DNA packaging process in the first study (#7, left) might involve homologous proteins as the DNA condensation process of the second study. Finally, both studies identify regulation of transcription (#5, left; #3, right) and transcription (#6 right, #2, left) as processes that utilize IDPs and proteins with IDRs.

As mentioned above, a weakness of the above computational approach for identifying IDP-associated biological processes is that most of the proteins containing IDRs also contain structured domains, so, for each of the indicated biological processes, it is unclear whether the underlying molecular functions depend on disorder or structure. Thus, follow-up studies are needed to confirm that IDRs are crucial for each biological process.

In the second study, the bioinformatics research was followed up by manual investigations of the literature to determine whether IDPs or IDRs actually played crucial roles in the IDP-associated biological processes. These literature investigations revealed at least one illustrative example of experimentally verified functional structure or disorder for a majority of the annotations that showed strong positive or negative correlation with predicted disorder *(7)*. Additional literature investigations were carried out to determine the extent to which IDPs or IDRs play key roles in the functions that provide the basis for differentiation. The literature reports provided experimental confirmation that IDPs or IDRs critically underlie the key molecular functions upon which differentiation depends. These functions include the following: cell adhesion, cell communication, 11 different developmental pathways crucial for metazoan differentiation, and gene regulation associated with body plan development for metazoans *(27)*.

To summarize, the molecular functions underlying these various IDP-associated biological processes often involve signaling and regulatory events that depend on reversible molecular interactions. Compared to structured proteins, IDPs evidently have advantages for carrying out the high-specificity, low-affinity interactions that underlie the reversibility needed for signaling *(28)*. In the context of the intrinsic disorder-based signaling, IDRs, posttranslational modifications (PTMs), and regions

encoded by alternatively spliced pre-mRNA (AS) were found to be colocalized for the important tumor suppressor p53 *(10)*. More recently, we showed that IDPs or IDRs, AS, and PTMs are found in the same protein for a large number of examples *(27, 29, 30)*, and for a few examples, these three protein features have been shown to collaborate and thereby to bring about highly complex signaling *(9)*.

Next, we offer several specific examples that provide experimentally backed support for the observations coming from the computational studies of the intrinsic disorder-driven biological processes.

1.3 Representative examples of IDPs and/or IDRs underlying various biological processes

1.3.1 Lac repressor

An IDR plays a crucial role in the function of the lac repressor, which was the first gene regulation system to be understood *(14)*; for developing this overall understanding, François Jacob and Jacques Monod were awarded the 1965 Nobel Prize in Physiology or Medicine. The lac repressor itself was not isolated and structurally characterized until later *(15, 16, 31, 32)*. The 347-residue lac repressor contains a 49-residue three-helix bundle DNA-recognition headpiece (residues 1–49), a 13-residue IDR connecting the headpiece to the core domain (residues 50–62), a 266-residue core domain (residues 63–329), and a four-helix bundle tetramerization domain, which is comprised of one 19-residue helix from the C-terminus of each lac repressor molecule *(16, 31)*. The lac repressor tetramer is organized as a pair of lac repressor dimers, with the two head pieces of each dimer binding to a single lac operon DNA segment. Thus, weak binding by a single headpiece becomes much stronger due to the avidity resulting from the binding of the second headpiece. The tetrameric lac repressor achieves cooperativity by binding simultaneously to the lac operon and to either one of two similar sequences located nearby, thereby forcing the intervening DNA into a loop *(33)*. The large separation of the two lac dimers in the tetramer helps to accommodate the looped DNA *(16, 31)*. The flexibility of the lac repressor IDR is essential for tight binding by the lac repressor dimer. That is, the high degree of flexibility of the 13-residue IDP linker enables the two head pieces to move independently, and thereby to separately undergo docking and to become simultaneously bound to their respective binding sites on the lac operon DNA.

Comparison of the structures of the free dimeric headpiece, the dimeric headpiece bound to nonspecific DNA (of the same length as the operon DNA), and the headpiece bound to operon DNA *(17)* highlights the role of the IDR in DNA recognition (Fig. 1.2). In the free form, each of the dimeric headpieces forms a three-helix bundle, but bundles remain highly dynamic by NMR, and the IDR remains unstructured. Upon binding to operon-length nonspecific DNA, the three-helix bundle exhibits essentially no reduction in its overall flexibility, and the IDR continues to be unstructured and highly mobile, both regions making multiple nonspecific electrostatic contacts

FIG. 1.2

Structural changes in the pathway of protein-DNA recognition by the lac repressor. The linker region (residues 50–62, *red* and *orange*) is an IDR in both the free state and when bound to nonspecific DNA. This IDR folds into an α-helix when bound into the minor groove of the natural operator *O1*. When the lac repressor binds to nonspecific DNA, the nucleic acid adopts a canonical B-DNA conformation. In contrast, when the lac repressor binds to its natural operator *O1*, the DNA becomes bent by ~36 degrees.

Adapted from Kalodimos, C. G.; Biris, N.; Bonvin, A. M.; Levandoski, M. M.; Guennuegues, M.; Boelens, R.; et al. Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. Science (New York, N.Y.) 2004, 305 (5682), 386–389.

with DNA. These multiple weak interactions primarily involving the DNA phosphates can exchange rapidly with different phosphates and by this means are thought to facilitate sliding along the DNA. Upon binding to lac-operon DNA, the two headpieces change their tilts with respect to the DNA and become tightly associated with the DNA, with a helix from each headpiece nestling into the major groove. This overall structural change is associated with specific hydrogen bonding interactions between several side chains located on the DNA-facing side of the helix and the bases comprising the DNA operon *(17)*. The IDP linkers also contribute to operon DNA recognition. The two linkers undergo disorder-to-order transitions, with residues 50–58 forming a helix that binds rather deeply into the minor groove. Minor-groove binding is facilitated by the intercalation of the side chains of L56 and L56' into the spaces between the stacked bases of CpG. This intercalation helps to pry apart the DNA bases, which in turn widens the minor groove, thus giving more space for helix binding and kinking the DNA. These studies suggest that the IDP linker in the lac repressor is not only important for DNA binding by facilitating simultaneous docking of pairs of headpieces, but that the IDP linker is also important for operon-sequence recognition and for helping to increase binding affinity.

The use of flexible IDP linkers to enable two DNA binding domains to associate simultaneously with two different DNA binding loci is also observed for pairs of eukaryotic transcription factors. For example, the two homeodomain-containing transcription factors, ultrabithorax and extradentical, simultaneously bind to their respective DNA binding loci *(8)*, thus enabling this pair of transcription factors to jointly regulate a number of different genes. These two molecules become bound to their respective loci while connected through a noncovalent interaction involving an ultrabithorax flexible IDP linker that utilizes a YPWM motif to bind to a specific site on the extradentical.

1.3.2 Nuclear factor of activated T-cells

Proteins involved in prokaryotic gene regulation such as the lac repressor have relatively small yet functionally important IDRs that typically comprise less than 30% of the protein *(34)*, whereas eukaryotic transcription factors contain much larger IDRs that typically comprise more than 50% of the protein *(34, 35)*. Recent studies show that putative intrinsic disorder is significantly enriched in eukaryotic DNA-binding proteins *(36)*. Moreover, as organisms become more complex (as estimated by their number of different cell types), the amount of predicted disorder shows a strongly correlated increase for certain transcription factors, namely those associated with cell cycle, cell size, cell division, cell differentiation, cell proliferation, and other important developmental processes *(37)*.

The nuclear factor of activated T-cells (NFAT) illustrates some of the roles of disorder in transcription factor function, namely modulation of function via molecular interactions, posttranslational modification, and alternative splicing. The nuclear factor of NFAT was discovered as an activator for the transcription of interleukin-2, which is a potent regulator of the immune response of T-cells *(38)*. NFAT signaling is widely used as a response regulator throughout the various cells of the immune system and in a number of other biological processes such as the inflammatory response, angiogenesis, cardiac valve formation, myocardial development, axonal guidance, skeletal muscle development, bone homeostasis, development and metastasis of cancer, and many other biological processes *(39)*.

The NFAT family has five members related by their similar DNA binding domain (DBD). The structured DBD and disordered regions of the five NFAT family members are shown in Fig. 1.3A along with additional data showing isoforms that arise from alternative splicing (Fig. 1.3B) and the structure/disorder and posttranslational modifications and other features of one member of the NFAT family (Fig. 1.3C). The structured DBD and IDR determined by prediction (Fig. 1.3A) agree reasonably well with CD and NMR data indicating the presence of a long IDR for one member of this protein family *(41)*. Note that the regions encoded by alternative splicing (AS) are mainly located in the IDR (Fig. 1.3B). The protein regions encoded by the alternatively spliced RNA of NFATc1 are all located in the disordered N- and/or C-terminus (Fig. 1.3B) and all the other NFAT genes have multiple AS isoforms at similar locations. AS is often tissue-specific and associated with cellular differentiation. Tissue-specific AS RNA encoded protein regions in NFAT have been suggested to contribute to altered transcriptional regulation in different types of T-cells *(42)*.

NFAT's IDR contains multiple segments used for signaling and regulation, including two CaN binding motifs, a nuclear export signal (NES), a nuclear localization signal (NLS), multiple sites for posttranslational modifications (PTMs) as well as other markers such as serine-rich regions (SRR) and serine-proline repeat motifs (SP) (Fig. 1.3C). The PTMs include phosphorylation of SRR1&2 and SP1–3 and sumoylation. These regulatory serine sites are phosphorylated by different kinases, specifically by PKA, DYRK, CK1, or GSK3 in a hierarchical pattern, creating a complex regulation that may allow for distinctive activation profiles in different

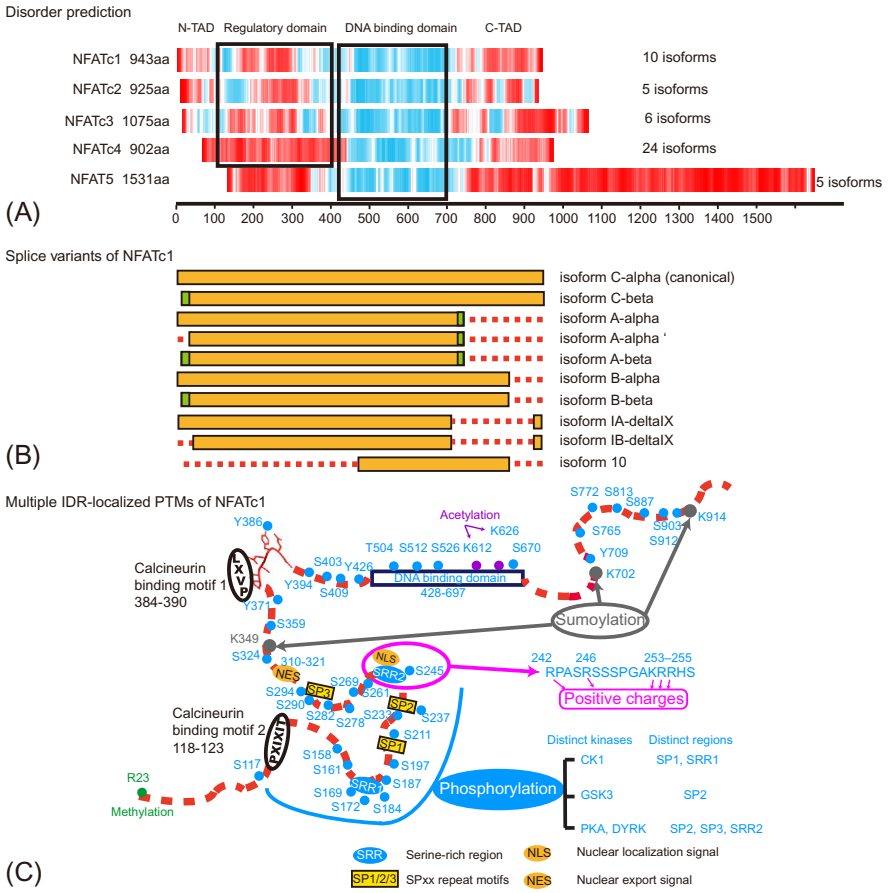

FIG. 1.3

Nuclear Factor of Activated T-cells (NFAT). (A) Disorder prediction of the five members of NFATs, where red indicates disorder and blue indicates structure. (B) The splice variants of NFATc1, where missing regions are indicated as a dashed line and replaced segments are in green. (C) Multiple PTMs from PhosphoSitePlus and available publications mainly localized within the IDR of NFATc1. PTMs are indicated by different colors. The important IDRs associated with localized functional regions are two calcineurin-binding motifs (PxIxIT and LxVP, x indicates any residue), a nuclear localization signal (NLS), three serine-proline-rich repeat motifs (SP1–3), and two serine-rich regions (SRR1&2). The structure of the second LxVP short motif is from the NFATc1-calcineurin binding complex (384–390 residues, the bound structure of which is shown in PDB id: 5SVE) *(40)*.

Reproduced with permission from Zhou J, Zhao S, Dunker AK. Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. J. Mol. Biol. 2018;430(16):2342–59.

cell types *(43)*. The sumoylation of NFATs, which is cell-specific and AS-isoform-specific, was recently shown to repress the transcriptional activity and regulate its nuclear retention, providing a new regulatory mechanism for NFAT functions *(44)*.

To summarize the regulatory events underlying the many biological processes listed above, following an appropriate stimulus, the levels of Ca^{2+} increase inside the cell, Ca^{2+} binds to calmodulin (CaM), and the Ca^{2+}/CaM complex binds to its target locus on CaN, thus activating CaN's S-T phosphatase activity. Next (or concurrently), one of NFAT's CaN binding motifs latches onto CaN. Because the phosphates and the binding motifs are colocalized in a common IDR on NFAT, the flexibility of this IDR allows the CaN to remove one phosphate after another while NFAT remains bound. When sufficient numbers of phosphates are removed, the NLS becomes active and NFAT translocates into the nucleus where it binds to DNA and activates a variety of genes that turn on cell division.

It is noteworthy that the carboxyl terminus of CaN ends in a long IDR that contains both an autoinhibitory domain and a Ca^{2+}/CaM binding domain. This long IDR in CaN thus provides the basis for CaN's activation at high Ca^{2+} and deactivation at low Ca^{2+} *(45)*. Also, CaM contains two Ca^{2+}-binding domains connected by a flexible IDP linker. This flexible linker enables CaM to wrap around its helical target as it binds *(46)*. Thus, overall, the signaling pathway involving CaM, CaN, and NFAT uses IDRs throughout to connect phosphorylation/dephosphorylation-based signaling with Ca^{2+}-based signaling. These are two of the most widely used signaling systems in eukaryotic cells.

1.3.3 p21, p27, and p57

Cell cycle regulation results from the inhibition of several different cyclin-dependent kinases (CDKs) when associated with their specific kinases. The well-studied CDK inhibitory proteins are p21 ($p21^{Waf1/Cip1/Sdi}$), p27 ($p27^{Kip1}$), and p57 ($p57^{Kip2}$) *(47)*. By associating with specific CDK-cyclin complexes, these proteins induce cell cycle arrest. When the production of these inhibitors is turned off and when the existing molecules are depleted by protease digestion, cell division resumes. The protease digestion of these proteins is not simple but instead depends on a highly regulated multistep process. Evidently, this regulated multistep process brings about the integration of different signals *(10, 13, 47, 48)*.

Studies based on NMR spectroscopy suggested the structural basis for cell-cycle regulation by one of these proteins, p21, and its binding to CDK2 *(48)*. That is, ^{15}N labeled p21 distinguished its resonances from those of the CDK2. Fig. 1.4 shows that, in the absence of CDK2, the ^{1}H-^{15}N heteronuclear single quantum correlation (HSQC) NMR spectrum of p21 is poorly dispersed and exhibits overlapping peaks. Only slight changes in this spectrum occur when 6 M urea is added, indicating the largely unstructured nature of free p21. When p21 is mixed with CDK2, the NMR spectrum becomes significantly dispersed (Fig. 1.4B), suggesting that p21 undergoes a disorder-to-order transition upon binding to CDK2. Also reported in 1996, the X-ray crystal structure of p27 bound to CDK2-Cyclin A shows how this rather

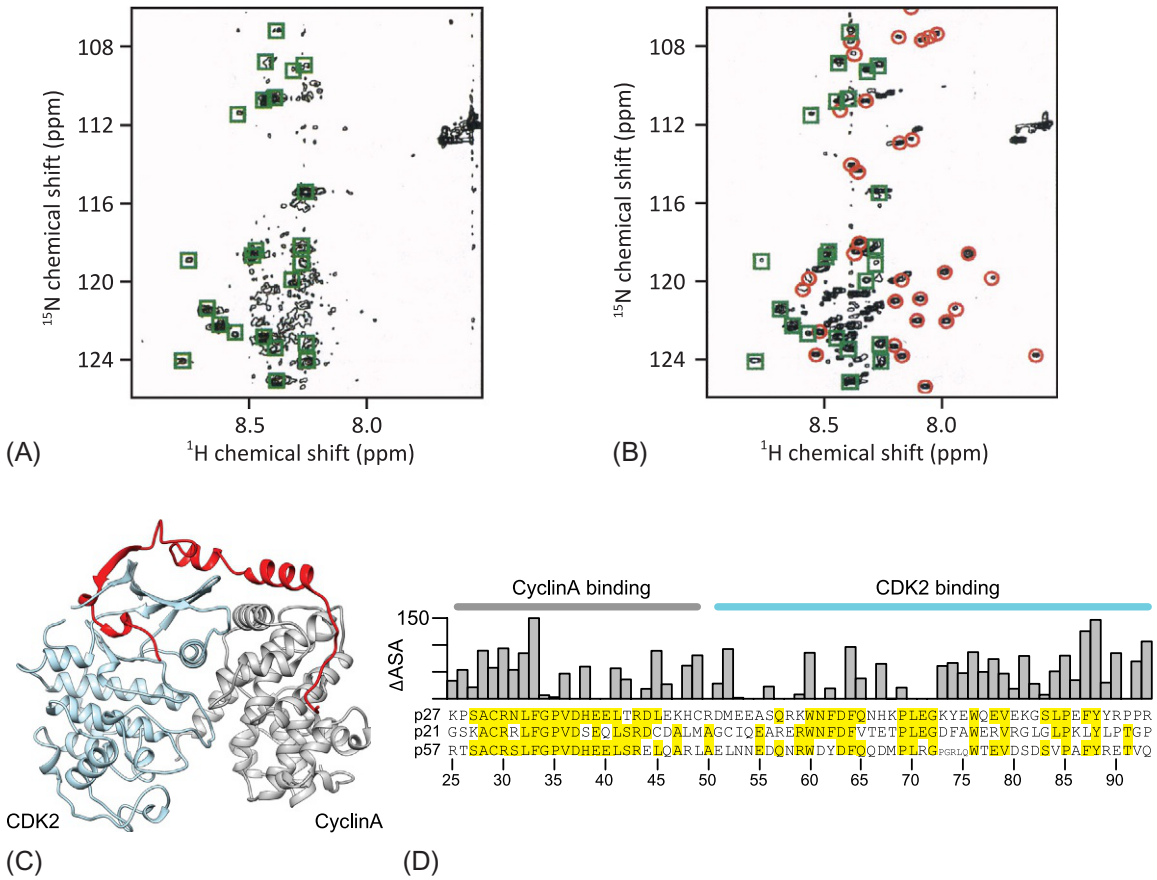

FIG. 1.4

Folding-and-binding in p21/p27 recognition of CDK-cyclin. The ^{1}H-^{15}N HSQC spectra of (A) free and (B) CDK2-bound p21 is shown with common (*green* boxes) and unique (*red* circles) resonances highlighted. (C) The structure (PDB ID 1JSU) of p27 (*red*) bound to CDK2 (*blue*) and cyclinA (*grey*). (D) Buried surface area (ΔASA) of p27 in the p27-CDK2/cyclinA complex and sequence conservation among p27, p21, and p57, where identical residues are highlighted in yellow.

(A) and (B) From Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. Structural Studies of p21Waf1/Cip1/Sdi1 in the Free and Cdk2-Bound State: Conformational Disorder Mediates Binding Diversity. Proc. Natl. Acad. Sci. U. S. A. 1996, 93 (21), 11504–11509; Copyright (1996) National Academy of Sciences, USA. (C) and (D) From Dunker, A. K.; Oldfield, C. J. Back to the Future: Nuclear Magnetic Resonance and Bioinformatics Studies on Intrinsically Disordered Proteins. In: Intrinsically Disordered Proteins Studied by NMR Spectroscopy; Felli, I. C., Pierattelli, R., Eds.; Springer International Publishing: Cham, 2015; pp 1–34. Originally modified from Russo, A. A.; Jeffrey, P. D.; Patten, A. K.; Massague, J.; Pavletich, N. P. Crystal Structure of the p27Kip1 Cyclin-Dependent-Kinase Inhibitor Bound to the Cyclin A-Cdk2 Complex. Nature 1996, 382 (6589), 325–331.

long IDP binds to both the kinase and its associated cyclin by wrapping around the heterodimeric complex (Fig. 1.4C) *(49)*. Finally, the binding region sequences of p21, p27, and p57 are aligned, showing their overall sequence similarity (Fig. 1.4D). Also shown in this panel is the buried surface estimated for p27 when it binds to the CDK2-Cyclin A complex. These data show that a fairly high sequence similarity occurs for the residues involved in the binding interface between p27 and the CDK2-Cyclin A complex, suggesting that p21 and p57 may bind similarly to their respective CDK-cyclin complexes.

Digestion of the p21, p27, or p57 molecule brings about continuation of the cell cycle progression. In the case of p27 bound to CDK2-cyclinA, this digestion likely involves conformational changes of the p27 that increase the accessibility of Y88. Given the elongated structure of p27 in the CDK2-cyclinA-p27 complex (Fig. 1.4C), perhaps some localized regions bind more weakly than others, thereby leading to localized separations of certain p27 sites along the binding interface. During such a transient exposure, p27's Y88 is phosphorylated by a nonreceptor tyrosine kinase. Once Y88 is phosphorylated, it remains constitutively separated from the remainder of the complex, thus making the CDK2 kinase active site more accessible. This facilitates the intracomplex phosphorylation of p27's T187. This second phosphorylation leads to ubiquitination, which is followed by the protease digestion of p27 by the proteasome. This multistep "signaling conduit" is speculated to provide a means to integrate multiple signals so that cell cycle progression proceeds only when the various signals have been received *(10–13)*. While still speculative at this time, the concept that multistep reversal of the very long interfaces of IDPs can bring about signal integration bears further investigation.

1.4 Biophysical features of IDPs and IDRs

Proteins are unbranched heteropolymers of 20 amino acids. It is thought that these natural polypeptides have originated as random copolymers of amino acids and were just evolutionary adjusted or "edited" (based on the principle of natural selection) to acquire and refine their various structures and functional properties *(51–54)*. Because of their very different physicochemical properties, the proteinaceous amino acid residues are engaged in different interactions of different natures (hydrophobic, electrostatic, or Van der Waals interactions and hydrogen bonding) with each other and with the protein's environment. This chemical diversity delimits the physicochemical heterogeneity of the resulting amino acid sequences and plays a crucial role in defining the structural heterogeneity of proteins. In fact, depending on the peculiarities of their amino acid sequences (or just their amino acid compositions) and environment, proteins either fold into unique structures or stay disordered to different degrees. Well before the concept of protein intrinsic disorder took its roots in modern protein science, it was recognized that information dictating unique structures of foldable proteins or protein domains is encoded in their amino acid sequences,. Therefore, at optimal conditions (temperature, solvent composition, pH, etc.), the native structure

represents a unique, stable, and kinetically accessible minimum of the free energy, a paradigm known as Anfinsen's dogma *(55)*. Therefore, the peculiarities of protein amino acid sequences define their ability to fold or to stay nonfolded (or intrinsically disordered) in specific environments. In agreement with this hypothesis, several systematic studies indicated that the absence of a rigid structure in IDPs and IDRs is encoded in the specific features of their amino acid sequences *(56–61)* that often contain numerous uncompensated charged groups *(62–64)* and possess a low content of hydrophobic amino acid residues *(62, 63)*.

An importance of the interplay between the amino acid sequence of a structured protein and its environment is illustrated by Fig. 1.5A, where a step-wise or framework model of protein folding is shown. According to this model, protein folds in such a way that the results of each step are not reconsidered, but just fastened at subsequent steps *(65–67)*. Here, the polypeptide chain first undergoes local folding to form the first folding intermediate (premolten globule) that contains some elements of secondary structure fluctuating around their native positions, which collapses next to form a second partially folded intermediate (molten globule). This molten globule is characterized by high compaction degree, has much more stable secondary structure and the main features of the native tertiary fold, but still does not have a unique three-dimensional (3D) structure that is formed at the last stage of the folding process

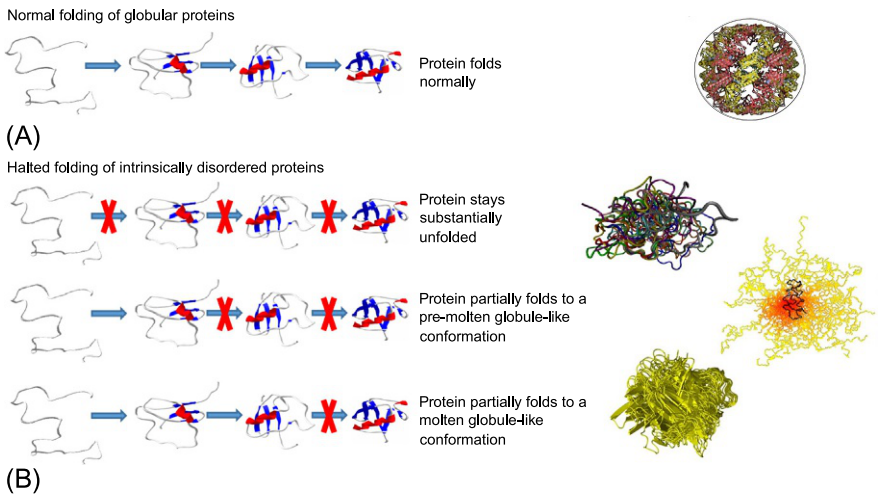

FIG. 1.5

Folding of a structured protein and halted folding of IDPs. (A) Framework model of a globular structured protein folding. Here, for a given foldable protein, different partially folded intermediates can be stabilized by changes in the environment. (B) Halted folding of IDPs, where the folding process seems to be stopped at different stages, thereby defining the different depth of disorder in a whole protein molecule. Here, differently underfolded species are observed at the same physiological conditions, and therefore the depth of disorder is defined by the peculiarities of an amino acid sequence of a given IDP.

(65–69). Both of these partially folded intermediates were observed as kinetic species that are transiently populated during the protein folding process. Furthermore, for several globular proteins, equilibrium premolten globule and molten globule intermediates were described at specific conditions (such as moderate concentrations of strong denaturing agents or extreme pH values). Therefore, for structured proteins, the existence of such partially folded intermediates under the equilibrium conditions is determined by the peculiarities of the protein environment that differently affect different conformational forces (that is, noncovalent interactions, such as hydrogen bonds, hydrophobic forces, electrostatic interactions, and van der Waals interactions, stabilizing the unique 3D structure of a globular protein). In fact, because conformational forces have different physicochemical natures, they differently respond to changes in the protein environment where some of the forces are either weakened or even completely eliminated under particular conditions, whereas other conformational forces remain unchanged or are even intensified under the same conditions. This gives rise to the formation of various partially folded conformations.

In contrast, IDPs can be disordered to different degrees (Fig. 1.5B). By the analogy to the partially folded intermediates of foldable proteins, IDPs can be classified as native coils, native premolten globules, and native molten globules, indicating that for these proteins, the folding process can be halted at different stages. Generally speaking, the structure of an IDP can be described as an ensemble of rapidly interchanging conformations, with some IDPs being more expanded (native coils and native premolten globules) and other IDPs being somewhat more compact (native molten globule) *(70)*. Because all these differently disordered conformations are present at the same (physiological) conditions, the different levels of disorder in these proteins are defined by the peculiarities of their amino acid sequences. One should keep in mind also that the aforementioned considerations are applicable not only to a whole protein or a protein domain, but also to different protein regions of various length. In other words, in contrast to the foldable single-domain proteins possessing a rather homogeneous distribution of the structure-coding potential within their sequences, IDPs and IDRs are characterized by the extremely high levels of sequence heterogeneity. This is because these sequences contain elements coding for potentially foldable, partially foldable, differently foldable, or not foldable at all protein segments that play different functional roles *(70, 71)*. For instance, some of the IDRs can bind to multiple partners gaining very different structures in the bound state *(72)*, and some undergo an order-to-disorder transition to make the protein active *(70, 73)*. Therefore, IDPs and IDRs can be described as a modular assembly of foldons, inducible foldons, semifoldons, nonfoldons, and unfoldons, and this mosaic organization of these proteins defines their extreme spatiotemporal heterogeneity related to their multifunctionality.

Another important biophysical feature of extended IDPs and IDRs (native coils and native premolten globules) is their specific "turned-out" response to changes in their environment, where they gain structure under conditions resulting in denaturation or even unfolding of ordered proteins *(70)*. For example, such extended IDPs can partially fold at elevated temperatures *(74–79)* due to the increased strength of

the hydrophobic interaction at higher temperatures *(74)*. Similarly, extended IDPs and IDRs gain more structure at extremely acidic and/or alkaline conditions *(74, 75, 80–83)*, being typically converted from a highly disordered coil-like conformation to a partially folded premolten globule-like conformation *(75, 80)*, due to the decrease in the charge-charge intramolecular repulsion that permits hydrophobic-driven collapse and formation of the partially folded conformations *(74)*. It is also important to note here that not only the full-length IDPs but also IDRs of different lengths can also possess such a "turned-out" response to changes in the protein's environment.

Finally, the lack of fixed structure in IDPs and IDRs possessing a wide spectrum of biological functions breaks a more than a century-old structure-function paradigm represented by the famous lock-and-key model. In this model, the specific function of a given protein is determined by its unique 3D structure, where protein and substrate have to fit to each other like a lock and key in order to exert a chemical effect on each other *(84)*. In fact, the inclusion of intrinsic disorder resulted in the conversion of the "one sequence-one structure-one function" model into a more realistic protein structure-function continuum representation of the correlation between protein structure and function, where a given protein exists as a dynamic conformational ensemble characterized by the diverse structural features and miscellaneous functions *(85)*. This structure-function continuum model also explains the important observation that the complexity of biological systems is mostly determined by their proteome sizes and not by the dimensions of their genomes *(86)*. In fact, the number of functionally different proteins found in eukaryotic organisms dramatically exceeds the number of genes. For instance, although the number of protein-coding genes in a human cell is approaching 20,700 *(87)*, the number of functionally different proteins is in a range of a few million. The functional diversification of proteinaceous products of a gene is achieved by single or multiple point mutations, indels, SNPs, alternative splicing, mRNA editing and other pretranslational mechanisms affecting mRNA as well as by a wide spectrum of various posttranslational modifications (PTMs) of a polypeptide chain.

All these mechanisms are responsible for the creation of a set of distinct protein molecules from a single gene, giving rise to the proteoform concept *(88)*. Recently, it was proposed that in addition to the aforementioned means that increase the chemical variability of a polypeptide chain, the protein structural diversity can be further increased by some other mechanisms, such as intrinsic disorder and function *(85)*. Because many PTM sites are known to be preferentially found within the IDRs *(89, 90)*, because mRNA regions affected by alternative splicing predominantly encode IDRs *(71)*, because IDPs and IDRs are known to be highly promiscuous binders *(10, 57, 60, 61, 70, 91–106)*, and because IDPs and IDRs are characterized by the exceptional spatiotemporal heterogeneity (see above), these proteins and protein regions represent a very rich source of proteoforms. Overall, proteoforms are classified as: 1) conformational or basic proteoforms generated due to the presence of intrinsic disorder in a protein; 2) functioning proteoforms generated by various perturbations induced in the conformational ensemble of an IDP/IDR by interaction with specific partners or any other function-related mechanisms; and 3) inducible or modified proteoforms generated by

allelic variations (that is, single- or multiple-point mutations, indels, SNPs), alternative splicing, and RNA editing of mRNA and numerous PTMs. All this defines the "protein structure-function continuum" concept, where a given protein exists as a dynamic conformational ensemble containing multiple proteoform types characterized by the diverse structural features and miscellaneous functions *(85)*.

1.5 Sequence features of IDPs and IDRs

The functional and biophysical properties of IDPs and IDRs are reflected in their sequences. The composition of IDRs is quite distinct from that of structured proteins. In addition, IDRs show patterns in sequence complexity and conservation that reflect the extremes of their molecular functions.

1.5.1 Compositional bias of IDPs

The inadequacy of IDRs to autonomously fold into a stable conformation is rooted in their amino acid composition. IDRs are characterized by a high net charge *(62–64)* and low hydrophobicity *(62, 63)*, relative to structured proteins. The relative compositional bias of IDRs, compared to the structured proteins (Fig. 1.6) demonstrates this bias. In general, IDRs are strongly depleted in aromatic residues, which often play key roles in the stabilization of protein structure *(110)*, large hydrophobic residues, and valine. IDRs are enriched in charged and hydrophilic residues. Glutamine and arginine are exceptions in this particular dataset, but they have also been found to be marginally biased toward IDRs in other datasets *(111)*; this depends on the details of the datasets. Arginine plays a nuanced role in IDRs. As a charged residue, it is often associated with IDRs because net charge is a strong determinant disorder *(61)*. It is also known to play a functional role in the interaction of some IDRs with RNA *(112)*. Conversely, it is known to stabilize protein structures through ready formation of salt bridges *(113)*.

The amino acids cysteine and proline play unique roles in protein structure, which is reflected in the extreme bias observed for these amino acids in IDPs. Under oxidizing conditions, cysteine forms covalent cross-links that stabilize protein structure. Additionally, cysteine may coordinate prosthetic groups, such as iron-sulfur centers, which also stabilize protein structure in addition to their functional aspects *(114)*. Under reducing conditions, cysteine cross-links are broken and stabilizing effects are lost, in many cases resulting in an inactive disordered form of these proteins *(115)*. This phenomenon has led to the idea of considering the redox environment when evaluating protein sequences for intrinsic disorder *(116)*.

In contrast to cysteine, proline tends to disrupt protein structure, and a relatively high proline composition is a common feature of IDRs. Proline restricts the conformation of protein backbone angles due to the pyrrolidine side chain, which prevents free rotation around the backbone C_α-N bond. Also, proline lacks a backbone amide hydrogen bond donor, which is required for stable secondary structure formation. These restrictions prevent proline residues from participating in the interiors

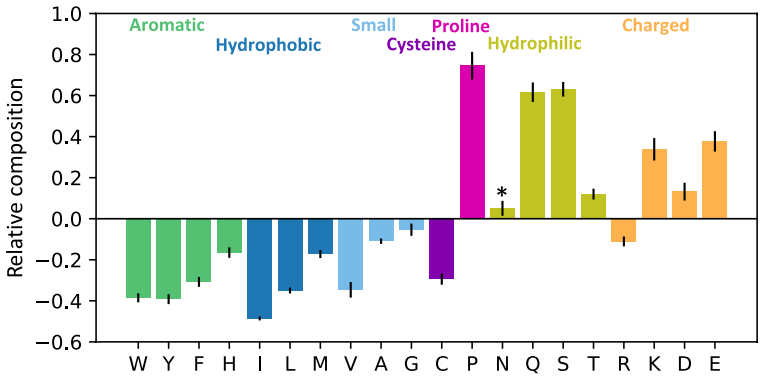

FIG. 1.6

Compositional bias of IDRs relative to structured proteins. Amino acids are grouped by their predominant property: aromatic, large hydrophobic, small hydrophobic, hydrophilic, and charged residues. Proline and cysteine are shown separately. IDRs are represented by all sequence unique IDRs from the DisProt database v7.5 *(107)*. Structured proteins are represented by structured domain sequences from the SCOP2 database *(108)*. The relative composition of each amino acid is calculated as its composition in IDRs minus its composition in ordered domains, divided by its composition in ordered domains, that is, positive values correspond to the enrichment in IDRs. Error bars indicate the 95% confidence interval derived from 10,000 bootstrap iterations *(109)*. All relative differences are significant (at an experimental error rate of 0.05), except for Asparagine (N), indicated with an asterisk.

of helices and sheets. Disruption of structure is the basis of proline's promotion of disorder in IDRs *(117)*. Somewhat paradoxically, poly-proline can form stable or pseudostable structures in the form of poly-proline II helices *(118)*. These structure share similar CD signatures with IDRs, leading to the suggestion that poly-proline II may be more common than is appreciated *(118)*.

Functional features of IDPs are also reflected by their composition. For instance, many transcription factors are known to be, or are likely to be, intrinsically disordered *(35)*. These proteins are often characterized by a high positive charge, which is advantageous when binding to negatively charged DNA. Functional regions of IDRs also often contradict the general compositional trends. For instance, the presence of aromatic residues within IDRs is often indicative of molecular recognition regions, and is a feature of IDR binding regions that may indicate interactions amenable to intervention *(119)*.

1.5.2 Sequence complexity and conservation

The diversity of amino acid types in a protein sequence is often related to the propensity for structured conformations. A common method to measure amino acid variety in a sequence is Shannon's entropy, also referred to as sequence complexity, where higher entropy values indicate a richer use of different amino acid types and lower entropy

values indicate use of fewer amino acid types. In general, globular proteins have a high sequence complexity, which is likely due to the requirement of forming a unique, stable structure *(111)*. On the other hand, fibrous proteins and coiled coils tend to have a lower sequence complexity due to their periodic sequences *(111)*. IDRs span a wide range of sequence complexity, from sequences as complex as the sequence of globular proteins down to sequences with as little complexity as fibrous proteins *(111)*.

Similar to sequence complexity, IDRs span the range from highly variable to highly conserved sequences. Many IDRs are highly variable *(120)* and susceptible to insertions and deletions across species *(121)*. Lack of conservation is likely related to reduced constraints for many IDR sequences, relative to structured proteins. That is because the structured proteins contain many intramolecular contacts and conservation of protein structure constrains the evolution of these sequences. However, the lack of conservation of a particular sequence does imply that there are no evolutionary constraints on intrinsically disordered proteins. Compositional constraints have been observed in several systems.

Conservation of IDR composition has been observed for several proteins where this conservation is vital for IDR function. For example, in histone H1, DNA binding and organizing activity has been found to rely only on protein composition rather than sequence; scrambling the sequence of this region has no impact on its DNA binding and organizing functions *(122)*. Similarly, the ubiquitin ligase protein san1 functions by binding misfolded proteins in a nonspecific manner. San1 homologs across closely related species show very little, if any, sequence conservation, but share a similar pattern of hydrophilic and hydrophobic regions, which are sufficient for conservation of function *(123)*.

However, the lack of conventional sequence conservation is not a defining feature of IDRs; many IDRs are as or more conserved than structured domains within the same protein *(120)*. Such conserved IDRs likely bind to partners with an accompanying disorder-to-order transition, where conservation arises from the need to recognize a partner specifically and form a specific structure in the complex. This relationship has been shown to hold for short linear motifs *(124)*. These short motif regions bind to specific partners and fold into specific structures, where these motifs are located within longer IDRs. Motif regions show greater sequence conservation than the surrounding IDR in general *(124)*. This phenomenon has also been observed in nonmotif binding regions. For example in the tumor suppressor p53, the N- and C-terminal intrinsically disordered tails show little sequence conservation except in the short binding regions located in the tails, where conservation is comparable to or greater than conservation of the DNA binding domain in this protein *(119)*.

1.6 Prediction of IDPs and IDRs

The presence of sequence biases that make the IDPs and IDRs distinct from the sequences of structured proteins and regions has motivated the development of computational models for the prediction of IDRs in protein sequences. These models

provide a score for each amino acid in the input protein sequence in the form of a real-value propensity for disordered conformation and/or a binary category (disordered versus ordered). The four-decade history of the field of computational disorder prediction is divided into three distinct periods that resulted in the development of the first-, second-, and third-generation methods *(125, 126)*.

The first-generation methods were developed between 1979 and 2001. Only several methods were released during that time, given a limited amount of IDP/IDR data that were available to develop and test these tools. The first computational approach that was published in 1979 was designed to predict random coil conformations *(127)*. However, it was later shown to suffer from a relatively poor predictive performance *(125)*. Two notable first-generation predictors were released by Romero, Obradovic, and Dunker in 1997 *(128)* and by Uversky and Fink in 2000 *(61)*.

The second-generation methods were made between 2002 and 2006. The defining characteristics of this period are a substantial increase in the number of the developed methods, the use of relatively simple predictive models, and the adoption of evolutionary profiles as inputs. The profiles are represented by the position specific score matrix (PSSM) that is typically generated using the PSI-BLAST algorithm *(129, 130)* from the input protein sequence. The acceleration of the development efforts coincided with the inclusion of the disorder prediction in the biannual CASP (Critical Assessment of protein Structure Prediction) experiments in 2002 *(131)*. The disorder predictions were assessed from CASP5 in 2002 until CASP10 in 2012 *(131–136)*. The second-generation tools predict IDRs based on two categories of models: 1) models produced with the help of machine-learning algorithms, with several popular examples that include PONDR predictors *(111, 137–141)*, DisEMBL *(142)*, DISOPRED *(143)*, RONN *(144)*, DISpro *(145)*, and NORSp *(146)*; and 2) *ab-initio* models that apply fundamental biophysical principles to make predictions directly from amino acid sequences, with examples being GlobPlot *(147)* and IUPred *(148, 149)*.

The third-generation methods were published after 2006. The essential features of these methods are the use of a sophisticated machine-learning model and the utilization of meta-predictor designs. Representative methods that take advantage of the complex machine-learning models include OnD-CRF, *(150)*, which uses the conditional random fields model; DNDisorder *(151)*, which applies a deep network and boosting; DISOPRED3 *(152)*, which combines three models: a support vector machine, a neural network, and the nearest neighbor; and SPOT-Disorder, which utilizes a deep recurrent neural network *(153)*. The meta-predictors combine outputs produced by several predictors, either via a majority vote consensus or a separate predictive model. Their underlying goal is to improve predictive performance when compared to the results produced by the input predictors. Examples of meta-predictors are MD *(154)*, CSpritz *(155)*, MetaDisorder *(156)*, disCoP *(157)*, MFDp *(158)*, metaPrDOS *(159)*, DisMeta *(160)*, CSpritz *(155)*, MFDp2 *(161, 162)*, and MobiDB-lite *(163)*. Additionally, a couple of methods rely on structural modeling to identify IDRs, such as PrDOS, which uses structural templates *(164)*, and DISOclust *(165, 166)*, which generates and processes multiple structural models.

Several dozens of disorder predictors have been developed so far. A fairly complete list of these methods can be compiled from several relevant surveys *(2, 125, 126, 167–172)*. Some of these tools are listed at the wiki page at https://en.wikipedia.org/wiki/List_of_disorder_prediction_software. Table 1.2 provides details for 10 popular disorder predictors that can be accessed and used online. Annual citations for these methods range between 16 and 115, attesting to the fact that they are frequently used.

These dozens of methods predict IDRs by relying on different datasets, different information computed from the input protein sequence, various predictive architectures, and a variety of model types. Correspondingly, their predictive performance also varies. One of the most popular measures of predictive quality is the area under the ROC curve (AUC). The AUC ranges between 0.5 (equivalent to random predictions) and 1 (always correct predictions). The AUC values for 30 methods that we published since 2004 are summarized in Fig. 1.7; we could not find AUC values for the earlier methods. These results were reported by the authors or listed in a comparative review in *(167)* for the methods for which the authors did not provide AUCs. These measurements were done using a range of datasets that have different levels of difficulty and that use different ways to define and annotate disorder, and thus they should not be directly compared between predictors. However, this analysis reveals an interesting long-term trend of gradually improving over time the predictive performance of the published tools. This is evidenced by the dotted line that correspond to a polynomial fit into the AUC values. The average AUC of the predictors published between 2004 and 2008 is 0.79. It went up to 0.85 over the following 5 years and it again trended up to 0.87 in the last 5 years. The newest methods, such as SPOT-Disorder-Single *(176)*, SPOT-Disorder *(153)*, AUCpreD *(177)*, and DeepCNF-D *(178)*, have reported AUCs > 0.88. This is in agreement with the best results reported in CASP10 where the AUC of the most accurate method has reached 0.91 *(135)*. This analysis suggests that the current methods offer very accurate predictions.

There are three databases that offer access to precomputed disorder predictions for large collections of proteins, as an alternative to making predictions with individual methods. They provide instantaneous access to predictions of several methods, without the need to run the predictions (which for some methods and longer protein sequences may take between several minutes and an hour) or traverse the Internet to find these predictors. However, they are limited to a certain set of proteins, outside of which the users will have to resort to using the predictors. These databases include the DICHOT database *(179)* that provides predictions from DISOPRED2 *(18)* and CLADIST *(180)*; the MobiDB database *(181–183)* that includes results generated with Espritz *(174)*, IUPred *(148)*, DisEMBL *(142)*, GlobPlot *(147)*, PONDR-VSL2b *(140)* and RONN *(144)*; and the D^2P^2 resource *(184)* that provides access to outputs from Espritz *(174)*, IUPred *(148)*, PV2 *(185)*, PrDOS *(164)*, PONDR-VSL2b *(140)*, and PONDR-VLXT *(111)*.

As discussed through this chapter, IDPs and proteins with IDRs are crucial for a wide range of cellular functions *(7, 19, 36, 91, 186–199)*. Correspondingly, recent years have seen a major shift in the development of computational tools, many of

Table 1.2 Popular disorder predictors

Name	References	Number of citations	Annual number of citations	URL
disCoP	*(157)*	87	21.7	http://biomine.cs.vcu.edu/servers/disCoP/
DisEMBL	*(142)*	1103	73.5	http://dis.embl.de/
DISOPRED3	*(143, 152, 173)*	604	43.1	http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1
ESpritz	*(174)*	186	31.0	http://protein.bio.unipd.it/espritz/
IUPred2	*(116, 148, 149)*	1502	115.5	https://iupred2a.elte.hu/
MFDp	*(158, 162)*	127	15.9	http://biomine.cs.vcu.edu/servers/MFDp
PONDR	*(137–141)*	426	32.8	http://www.pondr.com/ http://www.dabi.temple.edu/disprot/predictor.php
PrDOS	*(164)*	387	35.2	http://prdos.hgc.jp/
SPINE-D	*(175)*	110	18.3	http://sparks-lab.org/SPINE-D/
SPOT-Disorder	*(153)*	32	32	http://sparks-lab.org/server/SPOT-disorder/

Methods are listed in alphabetical order. Citation counts were collected from Google Scholar on Nov. 7, 2018, based on the most-cited publication for each tool, that is, one article was used for tools that were published multiple times to avoid duplicate citations. The annual number of citations is computed by dividing the number of citations by the number of years since the corresponding article was published.

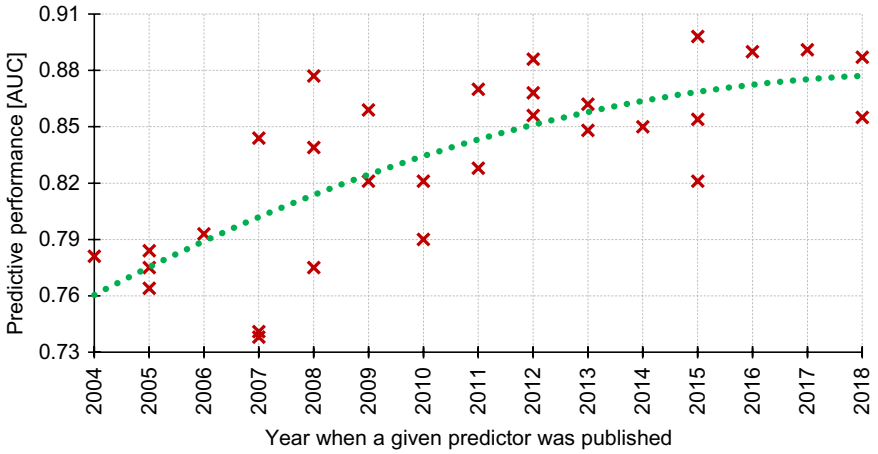

FIG. 1.7

Relation between predictive performance and the year of publication for 30 disorder predictors that were published since 2004. These methods include SPOT-Disorder-Single, IUpred2, SPOT-Disorder, AUCpreD, Disoclust2, DISOPRED3, DeepCNF-D, disCoP, DNDisorder, MFDp2, ESpritz, MetaDisorder, SPINE-D, CSpritz, IsUnstruct, MFDp, PONDR-FIT, MD, PreDisorder, Disoclust, metaPrDOS, OnD-CRF, NORSnet, Ucon, PrDOS, PONDR-VSL2B, DISpro, IUPred, RONN, and DISOPRED2.

which now target the prediction of specific functions of IDRs. These methods address mostly interaction-related functions that include protein-protein, protein-DNA, and protein-RNA binding. The largest group of predictors is dedicated to the prediction of protein-binding IDRs. They include alpha-MoRFpred *(200, 201)*, ANCHOR *(202, 203)*, retro-MoRF *(204)*, SLiMpred *(205)*, MoRFpred *(206, 207)*, PepBindPred *(208)*, MFSPSSMpred *(209)*, MoRFChiBi *(210)*, fMoRFpred *(211)*, disoRDPbind *(212, 213)*, DISOPRED3 *(152)*, MoRFChiBiWeb *(214)*, and ANCHOR2 *(116)*. Only one method, disoRDPbind *(212)*, is available to predict the DNA- and RNA-binding IDRs. The two newest tools target the prediction of multifunctional IDRs (DMRpred *(215)*) and disordered linkers (DFLpred *(216)*). This variety of recently developed tools empowers the community to not only accurately identify IDRs in protein sequences, but also to hypothesize about the functional roles of these regions.

1.7 Summary

IDPs and IDRs form ensembles of rapidly interchanging conformations and are disordered to different degrees, including native coils, native premolten globules, and native molten globules. They are abundant in nature as well as functionally important and distinct from their structured counterparts. Intrinsic disorder is indispensable for complex signaling and regulatory processes, which often depend on molecular interactions that involve proteins and nucleic acids. IDPs and IDRs are characterized

by unique biophysical properties and several sequence biases when compared to the structured proteins. These sequence-level biases motivated the development of computational methods that accurately predict IDRs and a selected set of functional types of IDRs from the protein sequence. While a lot has been learned about IDPs and IDRs in the last several decades, there are many outstanding mysteries surrounding this fascinating type of protein.

References

1. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114* (13), 6589–6631.
2. Lieutaud, P.; Ferron, F.; Uversky, A. V.; Kurgan, L.; Uversky, V. N.; Longhi, S. How Disordered Is My Protein and What Is Its Disorder for? A Guide Through the "Dark Side" of the Protein Universe. *Intrinsically Disord. Proteins* **2016**, *4* (1), e1259708.
3. Dunker, A. K.; Babu, M. M.; Barbar, E.; Blackledge, M.; Bondos, S. E.; Dosztányi, Z.; et al. What's in a Name? Why These Proteins Are Intrinsically Disordered. *Intrinsically Disord. Proteins* **2013**, *1* (1), e24157.
4. Manion, M. K.; O'Neill, J. W.; Giedt, C. D.; Kim, K. M.; Zhang, K. Y.; Hockenbery, D. M. Bcl-XL Mutations Suppress Cellular Sensitivity to Antimycin A. *J. Biol. Chem.* **2004**, *279* (3), 2159–2165.
5. Scalley-Kim, M.; Minard, P.; Baker, D. Low Free Energy Cost of Very Long Loop Insertions in Proteins. *Protein Sci.* **2003**, *12* (2), 197–206.
6. Mark, W. Y.; Liao, J. C.; Lu, Y.; Ayed, A.; Laister, R.; Szymczyna, B.; et al. Characterization of Segments From the Central Region of BRCA1: An Intrinsically Disordered Scaffold for Multiple Protein-Protein and Protein-DNA Interactions? *J. Mol. Biol.* **2005**, *345* (2), 275–287.
7. Xie, H.; Vucetic, S.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N.; et al. Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins With Long Disordered Regions. *J. Proteome Res.* **2007**, *6* (5), 1882–1898.
8. Passner, J. M.; Ryoo, H. D.; Shen, L.; Mann, R. S.; Aggarwal, A. K. Structure of a DNA-Bound Ultrabithorax-Extradenticle Homeodomain Complex. *Nature* **1999**, *397* (6721), 714–719.
9. Zhou, J.; Zhao, S.; Dunker, A. K. Intrinsically Disordered Proteins Link Alternative Splicing and Post-Translational Modifications to Complex Cell Signaling and Regulation. *J. Mol. Biol.* **2018**, *430* (16), 2342–2359.
10. Dunker, A. K.; Uversky, V. N. Signal Transduction via Unstructured Protein Conduits. *Nat. Chem. Biol.* **2008**, *4* (4), 229–230.
11. Follis, A. V.; Galea, C. A.; Kriwacki, R. W. Intrinsic Protein Flexibility in Regulation of Cell Proliferation: Advantages for Signaling and Opportunities for Novel Therapeutics. *Adv. Exp. Med. Biol.* **2012**, *725*, 27–49.
12. Galea, C. A.; Nourse, A.; Wang, Y.; Sivakolundu, S. G.; Heller, W. T.; Kriwacki, R. W. Role of Intrinsic Flexibility in Signal Transduction Mediated by the Cell Cycle Regulator, p27 Kip1. *J. Mol. Biol.* **2008**, *376* (3), 827–838.
13. Galea, C. A.; Wang, Y.; Sivakolundu, S. G.; Kriwacki, R. W. Regulation of Cell Division by Intrinsically Unstructured Proteins: Intrinsic Flexibility, Modularity, and Signaling Conduits. *Biochemistry* **2008**, *47* (29), 7598–7609.

14. Jacob, F.; Monod, J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* **1961**, *3*, 318–356.
15. Gilbert, W.; Muller-Hill, B. Isolation of the Lac Repressor. *Proc. Natl. Acad. Sci. U. S. A.* **1966**, *56* (6), 1891–1898.
16. Lewis, M.; Chang, G.; Horton, N. C.; Kercher, M. A.; Pace, H. C.; Schumacher, M. A.; et al. Crystal Structure of the Lactose Operon Repressor and Its Complexes With DNA and Inducer. *Science (New York, N.Y.)* **1996**, *271* (5253), 1247–1254.
17. Kalodimos, C. G.; Biris, N.; Bonvin, A. M.; Levandoski, M. M.; Guennuegues, M.; Boelens, R.; et al. Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science (New York, N.Y.)* **2004**, *305* (5682), 386–389.
18. Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins From the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337* (3), 635–645.
19. Peng, Z.; Yan, J.; Fan, X.; Mizianty, M. J.; Xue, B.; Wang, K.; et al. Exceptionally Abundant Exceptions: Comprehensive Characterization of Intrinsic Disorder in All Domains of Life. *Cell. Mol. Life Sci.* **2015**, *72* (1), 137–151.
20. Petrey, D.; Chen, T. S.; Deng, L.; Garzon, J. I.; Hwang, H.; Lasso, G.; et al. Template-Based Prediction of Protein Function. *Curr. Opin. Struct. Biol.* **2015**, *32*, 33–38.
21. Cruz, L. M.; Trefflich, S.; Weiss, V. A.; Castro, M. A. A. Protein Function Prediction. *Methods Mol. Biol.* **1654**, *2017*, 55–75.
22. Garfinkel, D. J. Genome Evolution Mediated by Ty Elements in Saccharomyces. *Cytogenet. Genome Res.* **2005**, *110* (1-4), 63–69.
23. Chant, J. Generation of Cell Polarity in Yeast. *Curr. Opin. Cell Biol.* **1996**, *8* (4), 557–565.
24. Madden, K.; Snyder, M. Cell Polarity and Morphogenesis in Budding Yeast. *Annu. Rev. Microbiol.* **1998**, *52*, 687–744.
25. Gancedo, J. M. Control of Pseudohyphae Formation in Saccharomyces cerevisiae. *FEMS Microbiol. Rev.* **2001**, *25* (1), 107–123.
26. Mackiewicz, P.; Kowalczuk, M.; Mackiewicz, D.; Nowicka, A.; Dudkiewicz, M.; Laszkiewicz, A.; et al. How Many Protein-Coding Genes Are There in the Saccharomyces cerevisiae Genome? *Yeast* **2002**, *19* (7), 619–629.
27. Dunker, A. K.; Bondos, S. E.; Huang, F.; Oldfield, C. J. Intrinsically Disordered Proteins and Multicellular Organisms. *Semin. Cell Dev. Biol.* **2015**, *37*, 44–55.
28. Schulz, G. E. Nucleotide Binding Proteins. In: Molecular Mechanism of Bilological Recognition; Balaban, M., Ed.; Elsevier/Noth-Holland Biomedical Press: New York, 1979; pp 79–84.
29. Niklas, K. J.; Bondos, S. E.; Dunker, A. K.; Newman, S. A. Rethinking Gene Regulatory Networks in Light of Alternative Splicing, Intrinsically Disordered Protein Domains, and Post-Translational Modifications. *Front. Cell Dev. Biol.* **2015**, *3*, 8.
30. Niklas, K. J.; Dunker, A. K.; Yruela, I. The Evolutionary Origins of Cell Type Diversification and the Role of Intrinsically Disordered Proteins. *J. Exp. Bot.* **2018**, *69* (7), 1437–1446.
31. Friedman, A. M.; Fischmann, T. O.; Steitz, T. A. Crystal Structure of Lac Repressor Core Tetramer and Its Implications for DNA Looping. *Science (New York, N.Y.)* **1995**, *268* (5218), 1721–1727.
32. Bell, C. E.; Lewis, M. Crystallographic Analysis of Lac Repressor Bound to Natural Operator O1. *J. Mol. Biol.* **2001**, *312* (5), 921–926.
33. Oehler, S.; Eismann, E. R.; Kramer, H.; Muller-Hill, B. The Three Operators of the Lac Operon Cooperate in Repression. *EMBO J.* **1990**, *9* (4), 973–979.

34. Minezaki, Y.; Homma, K.; Kinjo, A. R.; Nishikawa, K. Human Transcription Factors Contain a High Fraction of Intrinsically Disordered Regions Essential for Transcriptional Regulation. *J. Mol. Biol.* **2006**, *359* (4), 1137–1149.
35. Liu, J.; Perumal, N. B.; Oldfield, C. J.; Su, E. W.; Uversky, V. N.; Dunker, A. K. Intrinsic Disorder in Transcription Factors. *Biochemistry* **2006**, *45* (22), 6873–6888.
36. Wang, C.; Uversky, V. N.; Kurgan, L. Disordered Nucleiome: Abundance of Intrinsic Disorder in the DNA- and RNA-Binding Proteins in 1121 Species From Eukaryota, Bacteria and Archaea. *Proteomics* **2016**, *16* (10), 1486–1498.
37. Yruela, I.; Oldfield, C. J.; Niklas, K. J.; Dunker, A. K. Evidence for a Strong Correlation Between Transcription Factor Protein Disorder and Organismic Complexity. *Genome Biol. Evol.* **2017**, *9* (5), 1248–1265.
38. Shaw, J. P.; Utz, P. J.; Durand, D. B.; Toole, J. J.; Emmel, E. A.; Crabtree, G. R. Identification of a Putative Regulator of Early T Cell Activation Genes. *Science (New York, N.Y.)* **1988**, *241* (4862), 202–205.
39. Pan, M. G.; Xiong, Y.; Chen, F. NFAT Gene Family in Inflammation and Cancer. *Curr. Mol. Med.* **2013**, *13* (4), 543–554.
40. Sheftic, S. R.; Page, R.; Peti, W. Investigating the Human Calcineurin Interaction Network Using the piLxVP SLiM. *Sci. Rep.* **2016**, *6*, 38920.
41. Park, S.; Uesugi, M.; Verdine, G. L. A Second Calcineurin Binding Site on the NFAT Regulatory Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (13), 7130–7135.
42. Vihma, H.; Pruunsild, P.; Timmusk, T. Alternative Splicing and Expression of Human and Mouse NFAT Genes. *Genomics* **2008**, *92* (5), 279–291.
43. Taneri, B.; Snyder, B.; Novoradovsky, A.; Gaasterland, T. Alternative Splicing of Mouse Transcription Factors Affects Their DNA-Binding Domain Architecture and Is Tissue Specific. *Genome Biol.* **2004**, *5* (10), R75.
44. Hogan, P. G.; Chen, L.; Nardone, J.; Rao, A. Transcriptional Regulation by Calcium, Calcineurin, and NFAT. *Genes Dev.* **2003**, *17* (18), 2205–2232.
45. Rumi-Masante, J.; Rusinga, F. I.; Lester, T. E.; Dunlap, T. B.; Williams, T. D.; Dunker, A. K.; et al. Structural Basis for Activation of Calcineurin by Calmodulin. *J. Mol. Biol.* **2012**, *415* (2), 307–317.
46. Radivojac, P.; Vucetic, S.; O'Connor, T. R.; Uversky, V. N.; Obradovic, Z.; Dunker, A. K. Calmodulin Signaling: Analysis and Prediction of a Disorder-Dependent Molecular Recognition. *Proteins* **2006**, *63* (2), 398–410.
47. Lee, M. H.; Reynisdottir, I.; Massague, J. Cloning of p57KIP2, a Cyclin-Dependent Kinase Inhibitor With Unique Domain Structure and Tissue Distribution. *Genes Dev.* **1995**, *9* (6), 639–649.
48. Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. Structural Studies of p21Waf1/Cip1/Sdi1 in the Free and Cdk2-Bound State: Conformational Disorder Mediates Binding Diversity. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (21), 11504–11509.
49. Russo, A. A.; Jeffrey, P. D.; Patten, A. K.; Massague, J.; Pavletich, N. P. Crystal Structure of the p27Kip1 Cyclin-Dependent-Kinase Inhibitor Bound to the Cyclin A-Cdk2 Complex. *Nature* **1996**, *382* (6589), 325–331.
50. Dunker, A. K.; Oldfield, C. J. Back to the Future: Nuclear Magnetic Resonance and Bioinformatics Studies on Intrinsically Disordered Proteins. In: Intrinsically Disordered Proteins Studied by NMR Spectroscopy; Felli, I. C., Pierattelli, R., Eds.; Springer International Publishing: Cham, 2015; pp 1–34.
51. Lau, K. F.; Dill, K. A. Theory for Protein Mutability and Biogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87* (2), 638–642.

52. Ptitsyn, O. B. Molten Globule and Protein Folding. *Adv. Protein Chem.* **1995**, *47*, 83–229.
53. Ptitsyn, O. B.; Volkenstein, M. V. Protein Structure and Neutral Theory of Evolution. *J. Biomol. Struct. Dyn.* **1986**, *4* (1), 137–156.
54. Uversky, V. N. Protein Folding Revisited. A Polypeptide Chain at the Folding-Misfolding-Nonfolding Cross-Roads: Which Way to Go? *Cell. Mol. Life Sci.* **2003**, *60* (9), 1852–1871.
55. Anfinsen, C. B. Principles That Govern Folding of Protein Chains. *Science (New York, N.Y.)* **1973**, *181* (4096), 223–230.
56. Dunker, A. K.; Garner, E.; Guilliot, S.; Romero, P.; Albrecht, K.; Hart, J.; et al. Protein Disorder and the Evolution of Molecular Recognition: Theory, Predictions and Observations. *Pac. Symp. Biocomput.* **1998**, 473–484.
57. Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; et al. Intrinsically Disordered Protein. *J. Mol. Graph. Model.* **2001**, *19* (1), 26–59.
58. Uversky, V. N. Natively Unfolded Proteins: A Point Where Biology Waits for Physics. *Protein Sci.* **2002**, *11* (4), 739–756.
59. Uversky, V. N. What Does It Mean to Be Natively Unfolded? *Eur. J. Biochem.* **2002**, *269* (1), 2–12.
60. Uversky, V. N.; Dunker, A. K. Understanding Protein Nonfolding. *Biochim. Biophys. Acta* **2010**, *1804* (6), 1231–1264.
61. Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why Are "Natively Unfolded" Proteins Unstructured Under Physiologic Conditions? *Proteins* **2000**, *41* (3), 415–427.
62. Hemmings, H. C., Jr.; Nairn, A. C.; Aswad, D. W.; Greengard, P. DARPP-32, a Dopamine- and Adenosine 3':5'-Monophosphate-Regulated Phosphoprotein Enriched in Dopamine-Innervated Brain Regions. II. Purification and Characterization of the Phosphoprotein From Bovine Caudate Nucleus. *J. Neurosci.* **1984**, *4* (1), 99–110.
63. Gast, K.; Damaschun, H.; Eckert, K.; Schulze-Forster, K.; Maurer, H. R.; Muller-Frohne, M.; et al. Prothymosin Alpha: A Biologically Active Protein With Random Coil Conformation. *Biochemistry* **1995**, *34* (40), 13211–13218.
64. Weinreb, P. H.; Zhen, W.; Poon, A. W.; Conway, K. A.; Lansbury, P. T., Jr. NACP, a Protein Implicated in Alzheimer's Disease and Learning, Is Natively Unfolded. *Biochemistry* **1996**, *35* (43), 13709–13715.
65. Ptitsyn, O. B. Stages in the Mechanism of Self-Organization of Protein Molecules. *Dokl. Akad. Nauk SSSR* **1973**, *210* (5), 1213–1215.
66. Ptitsyn, O. B. Protein Folding: Hypothesis and Experiments. *J. Protein Chem.* **1987**, *6*, 273–293.
67. Ptitsyn, O. B.; Bychkova, V. E.; Uversky, V. N. Kinetic and Equilibrium Folding Intermediates. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **1995**, *348* (1323), 35–41.
68. Karplus, M.; Weaver, D. L. Protein-Folding Dynamics. *Nature* **1976**, *260* (5550), 404–406.
69. Kim, P. S.; Baldwin, R. L. Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. *Annu. Rev. Biochem.* **1982**, *51*, 459–489.
70. Uversky, V. N. Unusual Biophysics of Intrinsically Disordered Proteins. *Biochim. Biophys. Acta* **2013**, *1834* (5), 932–951.
71. Romero, P. R.; Zaidi, S.; Fang, Y. Y.; Uversky, V. N.; Radivojac, P.; Oldfield, C. J.; et al. Alternative Splicing in Concert With Protein Intrinsic Disorder Enables Increased Functional Diversity in Multicellular Organisms. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (22), 8390–8395.
72. Oldfield, C. J.; Meng, J.; Yang, J. Y.; Yang, M. Q.; Uversky, V. N.; Dunker, A. K. Flexible Nets: Disorder and Induced Fit in the Associations of p53 and 14-3-3 With Their Partners. *BMC Genomics* **2008**, *9* (Suppl 1), S1.

73. Jakob, U.; Kriwacki, R.; Uversky, V. N. Conditionally and Transiently Disordered Proteins: Awakening Cryptic Disorder to Regulate Protein Function. *Chem. Rev.* **2014**, *114* (13), 6779–6805.
74. Uversky, V. N. Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding. *Protein J.* **2009**, *28* (7-8), 305–325.
75. Uversky, V. N.; Li, J.; Fink, A. L. Evidence for a Partially Folded Intermediate in Alpha-Synuclein Fibril Formation. *J. Biol. Chem.* **2001**, *276* (14), 10737–10744.
76. Permyakov, S. E.; Millett, I. S.; Doniach, S.; Permyakov, E. A.; Uversky, V. N. Natively Unfolded C-Terminal Domain of Caldesmon Remains Substantially Unstructured After the Effective Binding to Calmodulin. *Proteins* **2003**, *53* (4), 855–862.
77. Uversky, V. N.; Permyakov, S. E.; Zagranichny, V. E.; Rodionov, I. L.; Fink, A. L.; Cherskaya, A. M.; et al. Effect of Zinc and Temperature on the Conformation of the Gamma Subunit of Retinal Phosphodiesterase: A Natively Unfolded Protein. *J. Proteome Res.* **2002**, *1* (2), 149–159.
78. Timm, D. E.; Vissavajjhala, P.; Ross, A. H.; Neet, K. E. Spectroscopic and Chemical Studies of the Interaction Between Nerve Growth Factor (NGF) and the Extracellular Domain of the Low Affinity NGF Receptor. *Protein Sci.* **1992**, *1* (8), 1023–1031.
79. Kim, T. D.; Ryu, H. J.; Cho, H. I.; Yang, C. H.; Kim, J. Thermal Behavior of Proteins: Heat-Resistant Proteins and Their Heat-Induced Secondary Structural Changes. *Biochemistry* **2000**, *39* (48), 14839–14846.
80. Uversky, V. N.; Gillespie, J. R.; Millett, I. S.; Khodyakova, A. V.; Vasiliev, A. M.; Chernovskaya, T. V.; et al. Natively Unfolded Human Prothymosin Alpha Adopts Partially Folded Collapsed Conformation at Acidic pH. *Biochemistry* **1999**, *38* (45), 15009–15016.
81. Konno, T.; Tanaka, N.; Kataoka, M.; Takano, E.; Maki, M. A Circular Dichroism Study of Preferential Hydration and Alcohol Effects on a Denatured Protein, Pig Calpastatin Domain I. *Biochim. Biophys. Acta* **1997**, *1342* (1), 73–82.
82. Lynn, A.; Chandra, S.; Malhotra, P.; Chauhan, V. S. Heme Binding and Polymerization by Plasmodium falciparum Histidine Rich Protein II: Influence of pH on Activity and Conformation. *FEBS Lett.* **1999**, *459* (2), 267–271.
83. Johansson, J.; Gudmundsson, G. H.; Rottenberg, M. E.; Berndt, K. D.; Agerberth, B. Conformation-Dependent Antibacterial Activity of the Naturally Occurring Human Peptide LL-37. *J. Biol. Chem.* **1998**, *273* (6), 3718–3724.
84. Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dt. Chem. Ges.* **1894**, *27*, 2985–2993.
85. Uversky, V. N. p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. *Int. J. Mol. Sci.* **2016**, *17* (11), 1874.
86. Schluter, H.; Apweiler, R.; Holzhutter, H. G.; Jungblut, P. R. Finding One's Way in Proteomics: a Protein Species Nomenclature. *Chem. Cent. J.* **2009**, *3*, 11.
87. Consortium, T. E. P. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **2012**, *489* (7414), 57–74.
88. Smith, L. M.; Kelleher, N. L. Consortium for Top Down P. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186–187.
89. Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; et al. The Importance of Intrinsic Disorder for Protein Phosphorylation. *Nucleic Acids Res.* **2004**, *32* (3), 1037–1049.
90. Pejaver, V.; Hsu, W. L.; Xin, F.; Dunker, A. K.; Uversky, V. N.; Radivojac, P. The Structural and Functional Signatures of Proteins That Undergo Multiple Events of Post-Translational Modification. *Protein Sci.* **2014**, *23* (8), 1077–1093.

91. Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **2002**, *41* (21), 6573–6582.
92. Dunker, A. K.; Brown, C. J.; Obradovic, Z. Identification and Functions of Usefully Disordered Proteins. *Adv. Protein Chem.* **2002**, *62*, 25–49.
93. Tompa, P. Intrinsically Unstructured Proteins. *Trends Biochem. Sci.* **2002**, *27* (10), 527–533.
94. Daughdrill, G. W.; Pielak, G. J.; Uversky, V. N.; Cortese, M. S.; Dunker, A. K. Natively Disordered Proteins. In: Handbook of Protein Folding; Buchner, J., Kiefhaber, T., Eds.; Wiley-VCH, Verlag GmbH & Co. KGaA: Weinheim, Germany, 2005; pp 271–353.
95. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Romero, P.; Uversky, V. N.; Dunker, A. K. Coupled Folding and Binding With Alpha-Helix-Forming Molecular Recognition Elements. *Biochemistry* **2005**, *44* (37), 12454–12470.
96. Radivojac, P.; Iakoucheva, L. M.; Oldfield, C. J.; Obradovic, Z.; Uversky, V. N.; Dunker, A. K. Intrinsic Disorder and Functional Proteomics. *Biophys. J.* **2007**, *92* (5), 1439–1456.
97. Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and Structure of Inherently Disordered Proteins. *Curr. Opin. Struct. Biol.* **2008**, *18* (6), 756–764.
98. Uversky, V. N. Multitude of Binding Modes Attainable by Intrinsically Disordered Proteins: A Portrait Gallery of Disorder-Based Complexes. *Chem. Soc. Rev.* **2011**, *40* (3), 1623–1634.
99. Uversky, V. N. Disordered Competitive Recruiter: Fast and Foldable. *J. Mol. Biol.* **2012**, *418* (5), 267–268.
100. Uversky, V. N.; Dunker, A. K. The Case for Intrinsically Disordered Proteins Playing Contributory Roles in Molecular Recognition Without a Stable 3D Structure. *F1000 Biol. Rep.* **2013**, *5*, 1.
101. Dyson, H. J.; Wright, P. E. Coupling of Folding and Binding for Unstructured Proteins. *Curr. Opin. Struct. Biol.* **2002**, *12* (1), 54–60.
102. Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*(3), 197–208.
103. Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* **1999**, *293* (2), 321–331.
104. Dunker, A. K.; Obradovic, Z. The Protein Trinity—Linking Function and Disorder. *Nat. Biotechnol.* **2001**, *19* (9), 805–806.
105. Mohan, A.; Oldfield, C. J.; Radivojac, P.; Vacic, V.; Cortese, M. S.; Dunker, A. K.; et al. Analysis of Molecular Recognition Features (MoRFs). *J. Mol. Biol.* **2006**, *362* (5), 1043–1059.
106. Vacic, V.; Oldfield, C. J.; Mohan, A.; Radivojac, P.; Cortese, M. S.; Uversky, V. N.; et al. Characterization of Molecular Recognition Features, MoRFs, and Their Binding Partners. *J. Proteome Res.* **2007**, *6* (6), 2351–2366.
107. Piovesan, D.; Tabaro, F.; Micetic, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; et al. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45* (D1), D219–d27.
108. Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A. G. SCOP2 Prototype: A New Approach to Protein Structure Mining. *Nucleic Acids Res.* **2014**, *42* (Database issue), D310–D314.
109. Vacic, V.; Uversky, V. N.; Dunker, A. K.; Lonardi, S. Composition Profiler: A Tool for Discovery and Visualization of Amino Acid Composition Differences. *BMC Bioinform.* **2007**, *8*, 211.

110. Burley, S.; Petsko, G. Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization. *Science (New York, N.Y.)* **1985**, *229* (4708), 23–28.
111. Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence Complexity of Disordered Protein. *Proteins* **2001**, *42* (1), 38–48.
112. Varadi, M.; Zsolyomi, F.; Guharoy, M.; Tompa, P. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One* **2015**, *10* (10), e0139731.
113. Siddiqui, K. S.; Cavicchioli, R. Cold-Adapted Enzymes. *Annu. Rev. Biochem.* **2006**, *75*, 403–433.
114. Kim, J. H.; Bothe, J. R.; Alderson, T. R.; Markley, J. L. Tangled Web of Interactions Among Proteins Involved in Iron-Sulfur Cluster Assembly as Unraveled by NMR, SAXS, Chemical Crosslinking, and Functional Studies. *Biochim. Biophys. Acta* **2015**, *1853* (6), 1416–1428.
115. Fraga, H.; Graña-Montes, R.; Illa, R.; Covaleda, G.; Ventura, S. Association Between Foldability and Aggregation Propensity in Small Disulfide-Rich Proteins. *Antioxid. Redox Signal.* **2014**, *21* (3), 368–383.
116. Meszaros, B.; Erdos, G.; Dosztanyi, Z. IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.* **2018**, *46* (W1), W329–W337.
117. Theillet, F. X.; Kalmar, L.; Tompa, P.; Han, K. H.; Selenko, P.; Dunker, A. K.; et al. The Alphabet of Intrinsic Disorder: I. Act Like a Pro: On the Abundance and Roles of Proline Residues in Intrinsically Disordered Proteins. *Intrinsically Disord. Proteins* **2013**, *1* (1), e24360.
118. Rath, A.; Davidson, A. R.; Deber, C. M. The Structure of "Unstructured" Regions in Peptides and Proteins: Role of the Polyproline II Helix in Protein Folding and Recognition. *Biopolymers* **2005**, *80* (2-3), 179–185.
119. Cheng, Y.; LeGall, T.; Oldfield, C. J.; Mueller, J. P.; Van, Y. Y.; Romero, P.; et al. Rational Drug Design via Intrinsically Disordered Protein. *Trends Biotechnol.* **2006**, *24* (10), 435–442.
120. Brown, C. J.; Johnson, A. K.; Dunker, A. K.; Daughdrill, G. W. Evolution and Disorder. *Curr. Opin. Struct. Biol.* **2011**, *21* (3), 441–446.
121. Light, S.; Sagit, R.; Sachenkova, O.; Ekman, D.; Elofsson, A. Protein Expansion Is Primarily Due to Indels in Intrinsically Disordered Regions. *Mol. Biol. Evol.* **2013**, *30* (12), 2645–2653.
122. Lu, X.; Hamkalo, B.; Parseghian, M. H.; Hansen, J. C. Chromatin Condensing Functions of the Linker Histone C-Terminal Domain Are Mediated by Specific Amino Acid Composition and Intrinsic Protein Disorder. *Biochemistry* **2009**, *48* (1), 164–172.
123. Boomsma, W.; Nielsen, S. V.; Lindorff-Larsen, K.; Hartmann-Petersen, R.; Ellgaard, L. Bioinformatics Analysis Identifies Several Intrinsically Disordered Human E3 Ubiquitin-Protein Ligases. *PeerJ.* **2016**, *4*, e1725.
124. Varadi, M.; Guharoy, M.; Zsolyomi, F.; Tompa, P. DisCons: A Novel Tool to Quantify and Classify Evolutionary Conservation of Intrinsic Protein Disorder. *BMC Bioinform.* **2015**, *16* (1), 153.
125. He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V. N.; Dunker, A. K. Predicting Intrinsic Disorder in Proteins: An Overview. *Cell Res.* **2009**, *19* (8), 929–949.
126. Meng, F.; Uversky, V. N.; Kurgan, L. Comprehensive Review of Methods for Prediction of Intrinsic Disorder and Its Molecular Functions. *Cell. Mol. Life Sci.* **2017**, *74* (17), 3069–3090.
127. RJP, W. The Conformation Properties of Proteins in Solution. *Biol. Rev.* **1979**, *54* (4), 389–437.

128. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E., Dunker, A. K., Eds.; Identifying Disordered Regions in Proteins From Amino Acid Sequence; Proceedings of International Conference on Neural Networks, Jun 9-12; 1997.
129. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; et al. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402.
130. Hu, G.; Kurgan, L. Sequence Similarity Searching. *Curr. Protoc. Protein Sci.* **2018**, e71.
131. Melamud, E.; Moult, J. Evaluation of Disorder Predictions in CASP5. *Proteins* **2003**, *53* (Suppl 6), 561–565.
132. Bordoli, L.; Kiefer, F.; Schwede, T. Assessment of Disorder Predictions in CASP7. *Proteins* **2007**, *69* (Suppl 8), 129–136.
133. Jin, Y.; Dunbrack, R. L., Jr. Assessment of Disorder Predictions in CASP6. *Proteins* **2005**, *61* (Suppl 7), 167–175.
134. Monastyrskyy, B.; Fidelis, K.; Moult, J.; Tramontano, A.; Kryshtafovych, A. Evaluation of Disorder Predictions in CASP9. *Proteins* **2011**, *79* (Suppl 10), 107–118.
135. Monastyrskyy, B.; Kryshtafovych, A.; Moult, J.; Tramontano, A.; Fidelis, K. Assessment of Protein Disorder Region Predictions in CASP10. *Proteins* **2014**, *82* (Suppl 2), 127–137.
136. Noivirt-Brik, O.; Prilusky, J.; Sussman, J. L. Assessment of Disorder Predictions in CASP8. *Proteins* **2009**, *77* (Suppl 9), 210–216.
137. Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. Flavors of Protein Disorder. *Proteins* **2003**, *52* (4), 573–584.
138. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Brown, C. J.; Dunker, A. K. Predicting Intrinsic Disorder From Amino Acid Sequence. *Proteins* **2003**, *53* (S6), 566–572.
139. Peng, K.; Vucetic, S.; Radivojac, P.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. Optimizing Long Intrinsic Disorder Predictors With Protein Evolutionary Information. *J. Bioinforma. Comput. Biol.* **2005**, *3* (1), 35–60.
140. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A. K. Exploiting Heterogeneous Sequence Properties Improves Prediction of Protein Disorder. *Proteins* **2005**, *61* (S7), 176–182.
141. Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A. K.; Obradovic, Z. Length-Dependent Prediction of Protein Intrinsic Disorder. *BMC Bioinform.* **2006**, *7* (1), 208.
142. Linding, R.; Jensen, L. J.; Diella, F.; Bork, P.; Gibson, T. J.; Russell, R. B. Protein Disorder Prediction: Implications for Structural Proteomics. *Structure* **2003**, *11* (11), 1453–1459.
143. Jones, D. T.; Ward, J. J. Prediction of Disordered Regions in Proteins From Position Specific Score Matrices. *Proteins* **2003**, *53* (S6), 573–578.
144. Yang, Z. R.; Thomson, R.; McNeil, P.; Esnouf, R. M. RONN: The Bio-Basis Function Neural Network Technique Applied to the Detection of Natively Disordered Regions in Proteins. *Bioinformatics* **2005**, *21* (16), 3369–3376.
145. Cheng, J.; Sweredoski, M.; Baldi, P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Min. Knowl. Disc.* **2005**, *11* (3), 213–222.
146. Liu, J.; Rost, B. NORSp: Predictions of Long Regions Without Regular Secondary Structure. *Nucleic Acids Res.* **2003**, *31*(13), 3833–3835.
147. Linding, R.; Russell, R. B.; Neduva, V.; Gibson, T. J. GlobPlot: Exploring Protein Sequences for Globularity and Disorder. *Nucleic Acids Res.* **2003**, *31* (13), 3701–3708.

148. Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: Web Server for the Prediction of Intrinsically Unstructured Regions of Proteins Based on Estimated Energy Content. *Bioinformatics* **2005**, *21* (16), 3433–3434.
149. Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated From Amino Acid Composition Discriminates Between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **2005**, *347* (4), 827–839.
150. Wang, L.; Sauer, U. H. OnD-CRF: Predicting Order and Disorder in Proteins Using [Corrected] Conditional Random Fields. *Bioinformatics* **2008**, *24* (11), 1401–1402.
151. Eickholt, J.; Cheng, J. DNdisorder: Predicting Protein Disorder Using Boosting and Deep Networks. *BMC Bioinform.* **2013**, *14* (1), 1–10.
152. Jones, D. T.; Cozzetto, D. DISOPRED3: Precise Disordered Region Predictions With Annotated Protein-Binding Activity. *Bioinformatics* **2015**, *31* (6), 857–863.
153. Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving Protein Disorder Prediction by Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. *Bioinformatics* **2017**, *33* (5), 685–692.
154. Schlessinger, A.; Punta, M.; Yachdav, G.; Kajan, L.; Rost, B. Improved Disorder Prediction by Combination of Orthogonal Approaches. *PLoS One* **2009**, *4* (2), e4433.
155. Walsh, I.; Martin, A. J. M.; Di Domenico, T.; Vullo, A.; Pollastri, G.; Tosatto, S. C. E. CSpritz: Accurate Prediction of Protein Disorder Segments With Annotation for Homology, Secondary Structure and Linear Motifs. *Nucleic Acids Res.* **2011**, *39* (Suppl 2), W190–W196.
156. Kozlowski, L. P.; Bujnicki, J. M. MetaDisorder: A Meta-Server for the Prediction of Intrinsic Disorder in Proteins. *BMC Bioinform.* **2012**, *13* (1), 1–11.
157. Fan, X.; Kurgan, L. Accurate Prediction of Disorder in Protein Chains With a Comprehensive and Empirically Designed Consensus. *J. Biomol. Struct. Dyn.* **2014**, *32* (3), 448–464.
158. Mizianty, M. J.; Stach, W.; Chen, K.; Kedarisetti, K. D.; Disfani, F. M.; Kurgan, L. Improved Sequence-Based Prediction of Disordered Regions With Multilayer Fusion of Multiple Information Sources. *Bioinformatics* **2010**, *26* (18), i489–i496.
159. Ishida, T.; Kinoshita, K. Prediction of Disordered Regions IN Proteins Based on the Meta Approach. *Bioinformatics* **2008**, *24* (11), 1344–1348.
160. Huang, Y. J.; Acton, T. B.; Montelione, G. T. DisMeta: A Meta Server for Construct Design and Optimization. *Methods Mol. Biol.* **2014**, *1091*, 3–16.
161. Mizianty, M. J.; Peng, Z.; Kurgan, L. MFDp2-Accurate Predictor of Disorder in Proteins by Fusion of Disorder Probabilities, Content and Profiles. *Intrinsically Disord. Proteins* **2013**, *1* (1), e24428.
162. Mizianty, M. J.; Uversky, V.; Kurgan, L. Prediction of Intrinsic Disorder in Proteins Using MFDp2. *Methods Mol. Biol.* **2014**, *1137*, 147–162.
163. Necci, M.; Piovesan, D.; Dosztanyi, Z.; Tosatto, S. C. E. MobiDB-Lite: Fast and Highly Specific Consensus Prediction of Intrinsic Disorder in Proteins. *Bioinformatics* **2017**, *33* (9), 1402–1404.
164. Ishida, T.; Kinoshita, K. PrDOS: Prediction of Disordered Protein Regions From Amino Acid Sequence. *Nucleic Acids Res.* **2007**, *35* (Suppl 2), W460–W464.
165. McGuffin, L. J. Intrinsic Disorder Prediction From the Analysis of Multiple Protein Fold Recognition Models. *Bioinformatics* **2008**, *24* (16), 1798–1804.
166. McGuffin, L. J.; Atkins, J. D.; Salehe, B. R.; Shuid, A. N.; Roche, D. B. IntFOLD: An Integrated Server for Modelling Protein Structures and Functions From Amino Acid Sequences. *Nucleic Acids Res.* **2015**, *43* (W1), W169–W173.
167. Peng, Z. L.; Kurgan, L. Comprehensive Comparative Assessment of In-Silico Predictors of Disordered Regions. *Curr. Protein Pept. Sci.* **2012**, *13* (1), 6–18.

168. Deng, X.; Eickholt, J.; Cheng, J. A Comprehensive Overview of Computational Protein Disorder Prediction Methods. *Mol. BioSyst.* **2012**, *8* (1), 114–121.
169. Walsh, I.; Giollo, M.; Di Domenico, T.; Ferrari, C.; Zimmermann, O.; Tosatto, S. C. E. Comprehensive Large-Scale Assessment of Intrinsic Protein Disorder. *Bioinformatics* **2015**, *31* (2), 201–208.
170. Meng, F.; Uversky, V.; Kurgan, L. Computational Prediction of Intrinsic Disorder in Proteins. *Curr. Protoc. Protein Sci.* **2017**, *88*, 2.16.1–2.16.14.
171. Uversky, V. N.; Radivojac, P.; Iakoucheva, L. M.; Obradovic, Z.; Dunker, A. K. Prediction of Intrinsic Disorder and Its Use in Functional Proteomics. *Methods Mol. Biol.* **2007**, *408*, 69–92.
172. Necci, M.; Piovesan, D.; Dosztanyi, Z.; Tompa, P.; Tosatto, S. C. E. A comprehensive Assessment of Long Intrinsic Protein Disorder From the DisProt Database. *Bioinformatics* **2018**, *34*, 445–452.
173. Ward, J. J.; McGuffin, L. J.; Bryson, K.; Buxton, B. F.; Jones, D. T. The DISOPRED Server for the Prediction of Protein Disorder. *Bioinformatics* **2004**, *20* (13), 2138–2139.
174. Walsh, I.; Martin, A. J. M.; Di Domenico, T.; Tosatto, S. C. E. ESpritz: Accurate and Fast Prediction of Protein Disorder. *Bioinformatics* **2012**, *28* (4), 503–509.
175. Zhang, T.; Faraggi, E.; Xue, B.; Dunker, A. K.; Uversky, V. N.; Zhou, Y. SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method. *J. Biomol. Struct. Dyn.* **2012**, *29* (4), 799–813.
176. Hanson, J.; Paliwal, K. K.; Zhou, Y. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. *J. Chem. Inf. Model.* **2018**.
177. Wang, S.; Ma, J.; Xu, J. AUCpreD: Proteome-Level Protein Disorder Prediction by AUC-Maximized Deep Convolutional Neural Fields. *Bioinformatics* **2016**, *32* (17), i672–i679.
178. Wang, S.; Weng, S.; Ma, J.; Tang, Q. DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields. *Int. J. Mol. Sci.* **2015**, *16* (8), 17315–17330.
179. Fukuchi, S.; Hosoda, K.; Homma, K.; Gojobori, T.; Nishikawa, K. Binary Classification of Protein Molecules Into Intrinsically Disordered and Ordered Segments. *BMC Struct. Biol.* **2011**, *11* (1), 1–10.
180. Fukuchi, S.; Homma, K.; Minezaki, Y.; Gojobori, T.; Nishikawa, K. Development of an Accurate Classification System of Proteins Into Structured and Unstructured Regions That Uncovers Novel Structural Domains: Its Application to Human Transcription Factors. *BMC Struct. Biol.* **2009**, *9* (1), 1–13.
181. Potenza, E.; Domenico, T. D.; Walsh, I.; Tosatto, S. C. E. MobiDB 2.0: An Improved Database of Intrinsically Disordered and Mobile Proteins. *Nucleic Acids Res.* **2015**, *43* (D1), D315–D320.
182. Di Domenico, T.; Walsh, I.; Martin, A. J. M.; Tosatto, S. C. E. MobiDB: A Comprehensive Database of Intrinsic Protein Disorder Annotations. *Bioinformatics* **2012**, *28* (15), 2080–2081.
183. Piovesan, D.; Tabaro, F.; Paladin, L.; Necci, M.; Micetic, I.; Camilloni, C.; et al. MobiDB 3.0: More Annotations for Intrinsic Disorder, Conformational Diversity and Interactions in Proteins. *Nucleic Acids Res.* **2018**, *46* (D1), D471–D476.
184. Oates, M. E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M. J.; Xue, B.; et al. D2P2: Database of Disordered Protein Predictions. *Nucleic Acids Res.* **2013**, *41* (D1), D508–D516.
185. Ghalwash, M. F.; Dunker, A. K.; Obradovic, Z. Uncertainty Analysis in Protein Disorder Prediction. *Mol. BioSyst.* **2012**, *8* (1), 381–391.

186. Babu, M. M. The Contribution of Intrinsically Disordered Regions to Protein Function, Cellular Complexity, and Human Disease. *Biochem. Soc. Trans.* **2016**, *44* (5), 1185–1200.
187. Xie, H.; Vucetic, S.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Obradovic, Z.; et al. Functional Anthology of Intrinsic Disorder. 3. Ligands, Post-Translational Modifications, and Diseases Associated With Intrinsically Disordered Proteins. *J. Proteome Res.* **2007**, *6* (5), 1917–1932.
188. Vucetic, S.; Xie, H.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Obradovic, Z.; et al. Functional Anthology of Intrinsic Disorder. 2. Cellular Components, Domains, Technical Terms, Developmental Processes, and Coding Sequence Diversities Correlated With Long Disordered Regions. *J. Proteome Res.* **2007**, *6* (5), 1899–1916.
189. Chowdhury, S.; Zhang, J.; Kurgan, L. In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome. *Proteomics* **2018**, e1800064.
190. Hu, G.; Wu, Z.; Uversky, V. N.; Kurgan, L. Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. *Int. J. Mol. Sci.* **2017**, *18* (12).
191. Yan, J.; Dunker, A. K.; Uversky, V. N.; Kurgan, L. Molecular Recognition Features (MoRFs) in Three Domains of Life. *Mol. BioSyst.* **2016**, *12* (3), 697–710.
192. Na, I.; Meng, F.; Kurgan, L.; Uversky, V. N. Autophagy-Related Intrinsically Disordered Proteins in Intra-Nuclear Compartments. *Mol. BioSyst.* **2016**, *12* (9), 2798–2817.
193. Meng, F.; Na, I.; Kurgan, L.; Uversky, V. N. Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments. *Int. J. Mol. Sci.* **2016**, *17* (1).
194. Xue, B.; Blocquel, D.; Habchi, J.; Uversky, A. V.; Kurgan, L.; Uversky, V. N.; et al. Structural Disorder in Viral Proteins. *Chem. Rev.* **2014**, *114* (13), 6880–6911.
195. Peng, Z.; Oldfield, C. J.; Xue, B.; Mizianty, M. J.; Dunker, A. K.; Kurgan, L.; et al. A Creature With a Hundred Waggly Tails: Intrinsically Disordered Proteins in the Ribosome. *Cell. Mol. Life Sci.* **2014**, *71* (8), 1477–1504.
196. Fan, X.; Xue, B.; Dolan, P. T.; LaCount, D. J.; Kurgan, L.; Uversky, V. N. The Intrinsic Disorder Status of the Human Hepatitis C Virus Proteome. *Mol. BioSyst.* **2014**, *10* (6), 1345–1363.
197. Peng, Z.; Xue, B.; Kurgan, L.; Uversky, V. N. Resilience of Death: Intrinsic Disorder in Proteins Involved in the Programmed Cell Death. *Cell Death Differ.* **2013**, *20* (9), 1257–1267.
198. Xue, B.; Mizianty, M. J.; Kurgan, L.; Uversky, V. N. Protein Intrinsic Disorder as a Flexible Armor and a Weapon of HIV-1. *Cell. Mol. Life Sci.* **2012**, *69* (8), 1211–1259.
199. Peng, Z.; Mizianty, M. J.; Xue, B.; Kurgan, L.; Uversky, V. N. More Than Just Tails: Intrinsic Disorder in Histone Proteins. *Mol. BioSyst.* **2012**, *8* (7), 1886–1901.
200. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K. Comparing and Combining Predictors of Mostly Disordered Proteins†. *Biochemistry* **2005**, *44* (6), 1989–2000.
201. Cheng, Y.; Oldfield, C. J.; Meng, J.; Romero, P.; Uversky, V. N.; Dunker, A. K. Mining α-Helix-Forming Molecular Recognition Features With Cross Species Sequence Alignments†. *Biochemistry* **2007**, *46* (47), 13468–13477.
202. Mészáros, B.; Simon, I.; Dosztányi, Z. Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Comput. Biol.* **2009**, *5* (5), e1000376.
203. Dosztányi, Z.; Mészáros, B.; Simon, I. ANCHOR: Web Server for Predicting Protein Binding Regions in Disordered Proteins. *Bioinformatics* **2009**, *25* (20), 2745–2746.

204. Xue, B.; Dunker, A. K.; Uversky, V. N. Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction. *Int. J. Mol. Sci.* **2010**, *11* (10), 3725–3747.
205. Mooney, C.; Pollastri, G.; Shields, D. C.; Haslam, N. J. Prediction of Short Linear Protein Binding Regions. *J. Mol. Biol.* **2012**, *415* (1), 193–204.
206. Disfani, F. M.; Hsu, W.-L.; Mizianty, M. J.; Oldfield, C. J.; Xue, B.; Dunker, A. K.; et al. MoRFpred, a Computational Tool for Sequence-Based Prediction and Characterization of Short Disorder-to-Order Transitioning Binding Regions in Proteins. *Bioinformatics* **2012**, *28* (12), i75–i83.
207. Oldfield, C. J.; Uversky, V. N.; Kurgan, L. Predicting Functions of Disordered Proteins With MoRFpred. *Methods Mol. Biol.* **2018**, *1851*.
208. Khan, W.; Duffy, F.; Pollastri, G.; Shields, D. C.; Mooney, C. Predicting Binding Within Disordered Protein Regions to Structurally Characterised Peptide-Binding Domains. *PLoS One* **2013**, *8* (9), e72838.
209. Fang, C.; Noguchi, T.; Tominaga, D.; Yamana, H. MFSPSSMpred: Identifying Short Disorder-to-Order Binding Regions in Disordered Proteins Based on Contextual Local Evolutionary Conservation. *BMC Bioinform.* **2013**, *14* (1), 1–14.
210. Malhis, N.; Gsponer, J. Computational Identification of MoRFs in Protein Sequences. *Bioinformatics* **2015**, *31*(11), 1738–1744.
211. Yan, J.; Dunker, A. K.; Uversky, V. N.; Kurgan, L. Molecular Recognition Features (MoRFs) in Three Domains of Life. *Mol. BioSyst.* **2015**.
212. Peng, Z.; Kurgan, L. High-Throughput Prediction of RNA, DNA and Protein Binding Regions Mediated by Intrinsic Disorder. *Nucleic Acids Res.* **2015**.
213. Peng, Z.; Wang, C.; Uversky, V. N.; Kurgan, L. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. *Methods Mol. Biol.* **2017**, *1484*, 187–203.
214. Malhis, N.; Jacobson, M.; Gsponer, J. MoRFchibi SYSTEM: Software Tools for the Identification of MoRFs in Protein Sequences. *Nucleic Acids Res.* **2016**.
215. Meng, F.; Kurgan, L. High-Throughput Prediction of Disordered Moonlighting Regions in Protein Sequences. *Proteins* **2018**, *86*, 1097–1110.
216. Meng, F.; Kurgan, L. DFLpred: High-Throughput Prediction of Disordered Flexible Linker Regions in Protein Sequences. *Bioinformatics* **2016**, *32* (12), i341–i350.