Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 568068, 14 pages
doi:10.1155/2010/568068

Review Article

The Mysterious Unfoldome: Structureless, Underappreciated, Yet Vital Part of Any Given Proteome

Vladimir N. Uversky[1, 2, 3]

[1] *Institute for Intrinsically Disordered Protein Research, The Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[2] *Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia*
[3] *Molecular Kinetics Inc., Indianapolis, IN 46268, USA*

Correspondence should be addressed to Vladimir N. Uversky, vuversky@iupui.edu

Received 22 July 2009; Accepted 10 September 2009

Academic Editor: Beatrix M. Ueberheide

Contrarily to the general believe, many biologically active proteins lack stable tertiary and/or secondary structure under physiological conditions in vitro. These intrinsically disordered proteins (IDPs) are highly abundant in nature and many of them are associated with various human diseases. The functional repertoire of IDPs complements the functions of ordered proteins. Since IDPs constitute a significant portion of any given proteome, they can be combined in an unfoldome; which is a portion of the proteome including all IDPs (also known as natively unfolded proteins, therefore, unfoldome), and describing their functions, structures, interactions, evolution, and so forth. Amino acid sequence and compositions of IDPs are very different from those of ordered proteins, making possible reliable identification of IDPs at the proteome level by various computational means. Furthermore, IDPs possess a number of unique structural properties and are characterized by a peculiar conformational behavior, including their high stability against low pH and high temperature and their structural indifference toward the unfolding by strong denaturants. These peculiarities were shown to be useful for elaboration of the experimental techniques for the large-scale identification of IDPs in various organisms. Some of the computational and experimental tools for the unfoldome discovery are discussed in this review.

1. Introducing Unfoldomes and Unfoldomics

Proteins are the major components of the living cell. They play crucial roles in the maintenance of life and protein dysfunctions may cause development of various pathological conditions. Although for a very long time it has been believed that the specific functionality of a given protein is predetermined by its unique 3D structure [1, 2], it is recognized now that the fate of any given polypeptide chain is determined by the peculiarities of its amino acid sequence. In fact, Figure 1 shows that although many proteins are indeed predisposed to fold into unique structures which evolved to possess unique biological functions, some proteins can misfold either spontaneously or due to the mutations and other genetic alterations, problematic processing or posttranslational modifications, or due to the exposure to harmful environmental conditions. Such misfolding is now considered as a crucial early step in the development of various protein conformation diseases [3]. Finally, for more than five decades, researchers have been discovering individual proteins that possess no definite ordered 3D structure but still play important biological roles. The discovery rate for such proteins has been increasing continually and has become especially rapid during the last decade [4]. Such proteins are widely known as intrinsically disordered proteins (IDPs) among other names and are characterized by the lack of a well-defined 3D structure under physiological conditions. The discovery and characterization of these proteins is becoming one of the fastest growing areas of protein science and it is recognized now that many such proteins with no unique structure have important biological functions [4–16]. Structural flexibility and plasticity originating from the lack of a definite-ordered 3D structure are believed to represent a major functional advantage for these proteins,

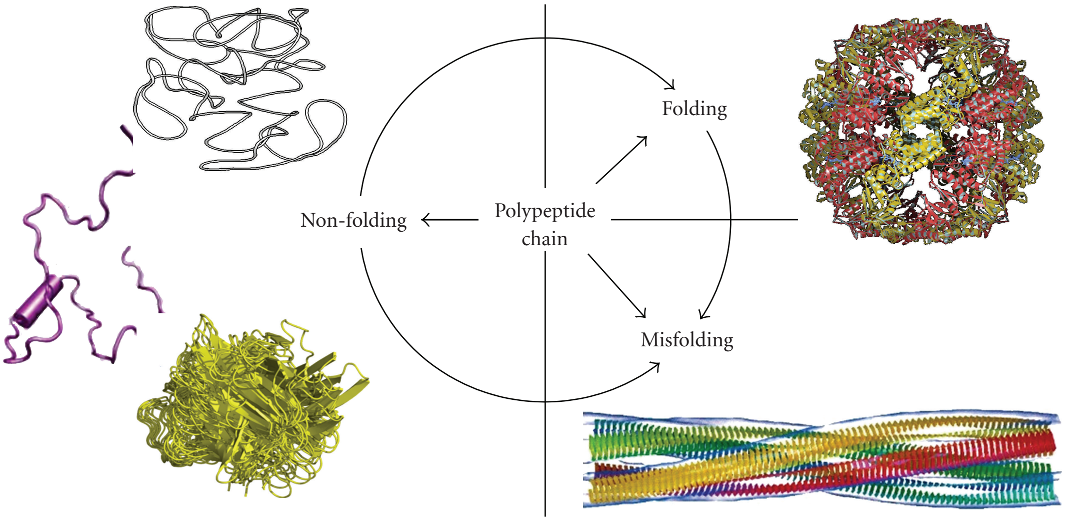


Figure 1: *Fate of a polypeptide chain. Left*. Three structures representing typical IDPs with different disorderedness levels (from top to bottom): native coil, native premolten globule, and native molten globule. *Right*. Top structure illustrates a well-folded protein, whereas the bottom structure represents one of the products of protein misfolding—a molecular model of the compact, 4-protofilament insulin fibril (http://people.cryst.bbk.ac.uk/~ubcg16z/amyloid/insmod.jpg).

enabling them to interact with a broad range of binding partners including other proteins, membranes, nucleic acids, and various small molecules [17–19].

The functions attributed to IDPs were grouped into four broad classes: (1) molecular recognition; (2) molecular assembly; (3) protein modification; and (4) entropic chain activities [5, 6]. IDPs are often involved in regulatory/signaling interactions with multiple partners that require high specificity and low affinity [7, 20]. Some illustrative biological activities of IDPs include regulation of cell division, transcription and translation, signal transduction, protein phosphorylation, storage of small molecules, chaperone action, and regulation of the self-assembly of large multiprotein complexes such as the ribosome [4–10, 13–16, 20–27]. The crucial role of IDPs in signaling is further confirmed by the fact that eukaryotic proteomes, with their extensively developed interaction networks, are highly enriched in IDPs, relative to bacteria and archaea (see below, [28–30]). Recently, application of a novel data mining tool to over 200 000 proteins from Swiss-Prot database revealed that many protein functions are associated with long disordered regions [13–15]. In fact, of the 711 Swiss-Prot functional keywords that were associated with at least 20 proteins, 262 were found to be strongly positively correlated with long intrinsically disordered regions (IDRs), whereas 302 were strongly negatively correlated with such regions [13–15]. Therefore, the functional diversity provided by disordered regions complements functions of ordered protein regions.

Although unbound IDPs are disordered in solution, they often perform their biological functions by binding to their specific partners. This binding involves a disorder-to-order transition in which IDPs adopt a highly structured conformation upon binding to their biological partners [31–38]. In this way, IDPs play diverse roles in regulating the function of their binding partners and in promoting the assembly of supramolecular complexes. Furthermore, because sites within their polypeptide chains are highly accessible, IDPs can undergo extensive posttranslational modifications, such as phosphorylation, acetylation, and/or ubiquitination (sumoylation, etc.), allowing for modulation of their biological activity or function. Intriguingly, IDPs were shown to be highly abundant in various diseases, giving rise to the "disorder in disorders" or D^2 concept which generally summarizes work in this area [39].

As the number of IDPs and IDRs in various proteomes is very large (e.g., for mammals, ~75% of their signaling proteins are predicted to contain long disordered regions (>30 residues), about half of their total proteins are predicted to contain such long disordered regions, and ~25% of their proteins are predicted to be fully disordered), and because IDPs and IDRs have amazing structural variability and possess a very wide variety of functions, the unfoldome and unfoldomics concepts were recently introduced [4, 40, 41]. The use of the suffix "-ome" has a long history while "-omics" is much more recent. The Oxford English Dictionary (OED) attributes "genome" to Hans Winkler from his 1920 work [42]. While the OED suggests that "genome" arose as a portmanteau of "gene" and "chromosome," this does not seem to be supported by literature. Instead, Lederberg and McCray suggest that as a botanist, Winkler must have been familiar with terms such as biome (a biological community), rhizome (a root system), and phyllome (the leaves covering a tree) among others, all of which were in use well before 1920 and all of which signify the collectivity of the units involved [43]. Thus, "ome" implies the complete set of the objects in question, with genome signifying the set of genes of an organism. By changing the "e" in "-ome" to "-ics," the new word is created that indicates the scientific study of the "-ome" in question. For example, officially the change of "genome" to "genomics" occurred in 1987, when a journal by this name was founded by Victor Lederberg and McCray [43].

Many additional conversions from –ome to –omics have subsequently occurred and a large number of "-omes" have been accepted in biology, including but not limited to the following: genome, proteome, interactome, metabolome, transcriptome, diseasome, toxicogenome, nutrigenome, cytome, oncoproteome, epitome, and glycome, and so forth. For a more complete list, the reader is directed to http://omics.org/.

Overall, the suffixes –ome and –omics imply a new layer of knowledge, especially when a scientist is dealing with the data produced by the large-scale studies, including the high-throughput experiments and the computational/bioinformatics analyses of the large datasets. The unfoldome and unfoldomics concepts are built on the ideas given above [44]. Unfoldome is attributed to a portion of proteome which includes a set of IDPs (also known as natively unfolded proteins, therefore, unfoldome). The term unfoldome is also used to cover segments or regions of proteins that remain unfolded in the functional state. Unfoldomics is the field that focuses on the unfoldome. It considers not only the identities of the set of proteins and protein regions in the unfoldome of a given organism but also their functions, structures, interactions, evolution, and so forth [44].

It is clearly recognized now that the disorderedness is linked to the peculiarities of amino acid sequences, as IDPs/IDRs exhibit low sequence complexity and are generally enriched in polar and charged residues and are depleted of hydrophobic residues (other than proline). These features are consistent with their inability to fold into globular structures and form the basis of computational tools for disorder prediction [8, 10, 45–48]. These same computational tools can also be utilized for the large-scale discovery of IDPs in various proteomes (see below).

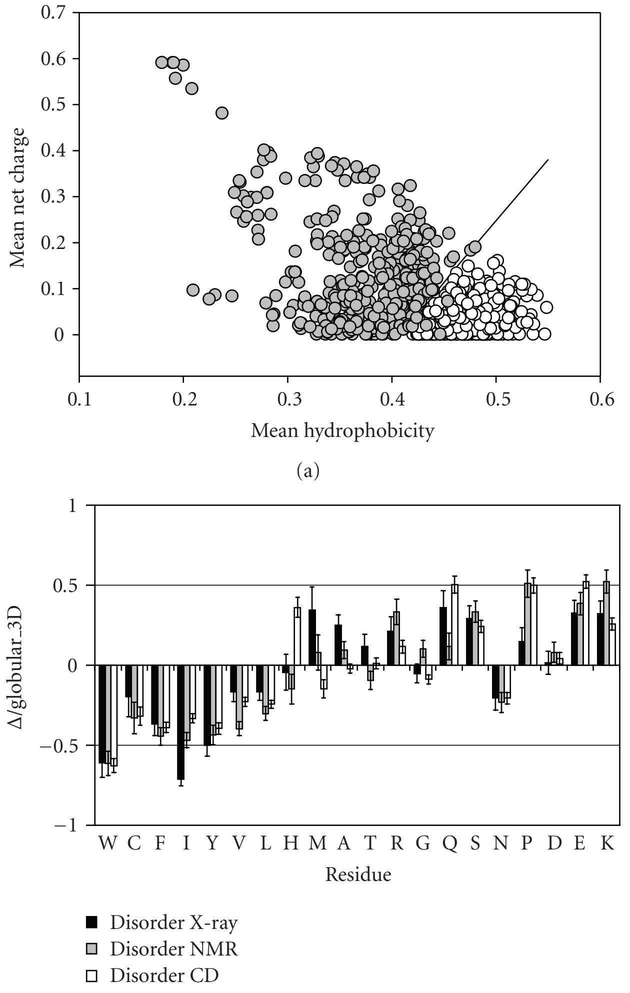
Being characterized by specific (and somewhat unique) amino acid sequences, IDPs possess a number of very distinctive structural properties that can be implemented for their discovery. This includes but is not limited to sensitivity to proteolysis [49], aberrant migration during SDS-PAGE [50], insensitivity to denaturing conditions [51], as well as definitive disorder characteristics visualized by CD spectropolarimetry, NMR spectroscopy, small-angle X-ray scattering, hydrodynamic measurement, fluorescence, as well as Raman and infrared spectroscopies [52, 53]. Structurally, intrinsically disordered proteins range from completely unstructured polypeptides to extended partially structured forms to compact disordered ensembles containing substantial secondary structure [4, 8, 9, 23, 54]. Many proteins contain mixtures of ordered and disordered regions. Extended IDPs are known to possess the atypical conformational behavior (such as "turn out" response to acidic pH and high temperature and insensitivity to high concentrations of strong denaturants), which is determined by the peculiarities of their amino acid sequences and the lack of ordered 3D structure [55]. These unique structural features of extended IDPs and their specific conformational behavior were shown to be useful in elaboration the experimental techniques for the large-scale identification of these important members of the protein kingdom. Three related methods were introduced: a method based on the finding that many proteins that fail to precipitate during perchloric acid or trichloroacetic acid treatment were IDPs [40]; a method utilizing the fact that IDPs possessed high resistance toward the aggregation induced by heat treatment [40, 56, 57]; and a method based on the heat treatment coupled with a novel 2D gel methodology to identify IDPs in cell extracts [56]. It is anticipated that these methodologies, combined with highly sensitive mass spectrometry-based techniques, can be used for the detection and functional characterization of IDPs in various proteomes. Some of the computational and experimental tools for the unfoldome discovery are discussed below in more details.

2. Computational Tools for Uncovering the Unfoldomes

2.1. Some Basic Principles of Disorder Prediction from Amino Acid Sequence. One of the key arguments about the existence and distinctiveness of IDPs came from various computational analyses. Historically, already at the early stages of the field, simple statistical comparisons of amino acid compositions and sequence complexity indicated that disordered and ordered regions are highly different to a significant degree [45, 58–60]. These sequence biases were then exploited to predict disordered regions or wholly disordered proteins with relatively high accuracy and to make crucial estimates about the commonness of disordered proteins in the three kingdoms of life [28, 45, 61].

Similar to the "normal" foldable proteins whose correct folding into the rigid biologically active conformation is determined by amino acid sequence, the absence of rigid structure in the "nontraditional" nonfoldable IDPs is encoded in the specific features of their amino acid sequences. In fact, some of the ID proteins have been discovered due their unusual amino acid sequence compositions and the absence of regular structure in these proteins has been explained by the specific features of their amino acid sequences including the presence of numerous uncompensated charged groups (often negative); that is, a large net charge at neutral pH, arising from the extreme pI values in such proteins [62–64], and a low content of hydrophobic amino acid residues [62, 64]. Interestingly, the first predictor of intrinsic disorder was developed by R.J.P. Williams based on the abnormally high charge/hydrophobic ratio for the two ID proteins; that is, using the same set of attributes, large net charge and low overall hydrophobicity [65]. Although this predictor was used to separate just two ID proteins from a small set of ordered proteins, this paper is significant as being the first indication that ID proteins have amino acid compositions that differ substantially from those of proteins with 3D structure.

Later, this approach was re-invented in a form of charge-hydropathy plot [45]. To this end, 275 natively folded and 91 natively unfolded proteins (i.e., proteins which at physiologic conditions have been reported to have the NMR chemical shifts of a random-coil, and/or lack significant ordered secondary structure (as determined by CD or FTIR), and/or show hydrodynamic dimensions close to those typical of an unfolded polypeptide chain) have been assembled from the literature searches. From the comparison of these datasets it has been concluded that the combination of low mean hydrophobicity and relatively high net charge represents an important prerequisite for the absence of compact structure in proteins under physiological conditions. This observation was used to develop a charge-hydropathy (CH) plot method of analysis that distinguishes ordered and disordered proteins based only on their net charges and hydropathy [45]. Figure 2(a) represents the original CH-plot and shows that natively unfolded proteins are specifically localized within a unique region of CH phase space. Furthermore, ID and ordered proteins can be separated by a linear boundary, above which a polypeptide chain with a given mean net charge will most probably be unfolded [45]. From the physical viewpoint, such a combination of low hydrophobicity with high net charge as a prerequisite for intrinsic unfoldedness makes perfect sense: high net charge leads to charge-charge repulsion, and low hydrophobicity means less driving force for protein compaction. In other words, these features are characteristic for ID proteins with the coil-like (or close to coil-like) structures. Obviously, such highly disordered proteins represent only a small subset of the ID protein realm.



(a)

(b)

Figure 2: *Peculiarities of amino acid composition of ID proteins.* (a) Comparison of the mean net charge and the mean hydrophobicity for a set of 275 ordered (open circles) and 91 natively unfolded proteins (gray circles). The solid line represents the border between extended IDPs and ordered proteins (see text). (b) Order/disorder composition profile. Comparisons of amino acid compositions of ordered proteins with each of three databases of disordered proteins. The ordinates are (%amino acid in disordered dataset – %amino acid in ordered dataset)/(%amino acid in ordered dataset) = △/globular_3D. Names of each database indicate how the disordered regions were identified. Negative values indicate that the disordered database has less than order, positive indicates more than order.

More detailed analysis was elaborated to gain additional information on the compositional difference between ordered and ID proteins. Comparison of a nonredundant set of ordered proteins with several datasets of disorder (where proteins were grouped based on different techniques, X-ray crystallography, NMR and CD, used to identify disorder) revealed that disordered regions share at least some common sequence features over many proteins [66, 67]. These differences in amino acid compositions are visualized in Figure 2(b). Here, the relative content of each amino acid in a given disordered dataset has been expressed as (Disordered–Ordered)/(Ordered). Thus, negative peaks correspond the amino acids in which the disordered segments are depleted compared with the ordered ones, and positive peaks indicate the amino acids in which ID regions are enriched [8]. The arrangement of the amino acids from least to most flexible was based on the scale established by Vihinen et al. [68]. This scale was defined by the average residue B-factors of the backbone atoms for 92 unrelated proteins. Figure 2(b) shows that the disordered proteins are significantly depleted in bulky hydrophobic (Ile, Leu, and Val) and aromatic amino acid residues (Trp, Tyr, and Phe), which would normally form the hydrophobic core of a folded globular protein, and also possess low content of Cys and Asn residues. The depletion of ID protein in Cys is also crucial as this amino acid residue is known to have a significant contribution to the protein conformation stability via the disulfide bond formation or being involved in coordination of different prosthetic groups. These depleted residues, Trp, Tyr, Phe, Ile, Leu, Val, Cys, and Asn were proposed to be called order-promoting amino acids. On the other hand, ID proteins were shown to be substantially enriched in polar, disorder-promoting, amino acids: Ala, Arg, Gly, Gln, Ser, Glu, and Lys and also in the hydrophobic, but structure-braking Pro [8, 69, 70]. Note that these biases in the amino acid compositions of ID proteins are also consistent with the low overall hydrophobicity and high net charge characteristic of the natively unfolded proteins [45].

In addition to amino-acid composition, the disordered segments have also been compared with the ordered ones by various attributes such as hydropathy, net charge, flexibility index, helix propensities, strand propensities, and compositions for groups of amino acids such as W + Y + F (aromaticity). As a result, 265 property-based attribute scales [69] and more than 6000 composition-based attributes (e.g., all possible combinations having one to four amino acids in the group) have been compared [71]. It has been established that ten of these attributes, including 14 Å contact number, hydropathy, flexibility, β-sheet propensity, coordination number, R+E+S+P, bulkiness, C+F+Y+W, volume, and net charge, provide fairly good discrimination between order and disorder [8]. Later, it has been shown that not only the sequence compositions of ordered and disordered regions were different but also that disordered regions of various lengths were diverse as well. In particular, four classes of protein regions were compared: (a) low B-factor ordered regions; (b) high B-factor ordered regions; (c) short disordered regions; and (d) long disordered regions [72]. The four types of regions were shown to have distinct sequence and physicochemical characteristics, with short disordered regions and high B-factor regions being the two closest groups. Furthermore, each of these two groups was closer to the long disordered regions than to the rigid ordered regions. In summary, the analysis of sequence and comparison of their various physicochemical properties indicated that all sets were mutually different [72]. For example, the short disordered and high B-factor regions were shown to be more negatively charged, while long disordered regions were either positively or negatively charged, but on average nearly neutral [72].

As the amino acid sequences of the IDPs and IDRs differ dramatically from those of the ordered proteins and regions, these amino acid sequence differences were used to develop various predictors of intrinsic disorder. As it has been already mentioned, based on a very small number of proteins, Williams suggested an approach for using amino acid sequence for identifying proteins that form random coils rather than globular structures [65], but this approach was never carefully tested. Later, the first well-tested predictors of IDPs were independently published [45, 58].

Protein disorder is a multifaced phenomenon; that is, disordered proteins, being mobile, flexible, and dynamic, might have very different structural features, which range from collapsed molten globule-like conformation to extended coil-like state. It has been suggested that just as an ordered protein is comprised of different types of secondary structure (α-helices, β-strands, β-turns, 3_{10}-helices, and others), ID protein can also be made up of distinguishable types of disorder [73]. To check this hypothesis, a partitioning algorithm based on the differential prediction accuracies has been developed [73]. This algorithm used the notion that a specialized predictor built on a given disorder flavor should have significantly higher same-flavor accuracy than other-flavor predictors or than a global predictor applied to the same given flavor. Application of this partitioning algorithm to known disordered proteins identified three distinctive “flavors” of disorder, arbitrarily called V, C, and S [73]. Importantly, the flavor-specific disordered proteins have been shown to be distinguishable not only by their amino acid compositions but also by disordered sequence locations, and biological functions. Based on these observations, it was proposed that specific flavor-function relationships do exist and thus it is possible (in principle) to identify the functions of disordered regions from their amino acid sequences alone, without any need for specific structural knowledge [73].

Since then, numerous researchers have designed many algorithms to predict disordered proteins utilizing specific biochemical properties and biased amino acid compositions of IDPs. Various prediction ideas and different computing techniques have been utilized. Many of these predictors, including PONDRs [58, 70, 74–77], FoldIndex [78], GlobPlot [79], DisEMBL [80], DISOPRED and DISOPRED2 [29, 81–83], IUPred [84], FoldUnfold [85], RONN [86], DisPSSMP [87], DisPSSMP2 [88], Spritz [89], and PrDOS [90], and so forth, can be accessed via public servers and evaluate intrinsic disorder on a per-residue basis. Since the first predictors were published, more than 50 predictors of disorder have been developed [91].

It is important to remember that comparing and combining several predictors on an individual protein of interest or on a protein dataset can provide additional insight regarding the predicted disorder if any exists. This is illustrated by a study where two distinct methods for using amino acid sequences to predict which proteins are likely to be mostly disordered, cumulative distribution function (CDF) analysis and charge-hydropathy (CH) plot, have been compared [30]. CDF is based on the PONDR VLXT predictor, which predicts the order-disorder class for every residue in a protein [28, 30]. CDF curves for PONDR VLXT predictions begin at the point with coordinates (0,0) and end at the point with coordinates (1,1) because PONDR VLXT predictions are defined only in the range (0,1) with values less than 0.5 indicating a propensity for order and values greater than or equal to 0.5 indicating a propensity for disorder. The optimal boundary that provided the most accurate order-disorder classification was determined and it has been shown that seven boundary points located in the 12th through 18th bin provided the optimal separation of the ordered and disordered protein sets [30]. For CDF analysis, order-disorder classification is based on whether a CDF curve is above or below a majority of boundary points [30]. In summary, CDF analysis summarizes the per-residue predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores. The other method of order-disorder classification is charge-hydropathy plots [45], in which ordered and disordered proteins being plotted in charge-hydropathy space can be separated to a significant degree by a linear boundary. It has been established that CDF analysis predicts a much higher frequency of disorder in sequence databases than CH-plot discrimination [30]. However, the vast majority of disordered proteins predicted by charge-hydropathy discrimination were also predicted by CDF analysis. These findings are not a big surprise, as CH-plot analysis discriminates protein using only two attributes, mean net charge and mean hydrophobicity, whereas PONDR VLXT (and consequently CDF) is a neural network, which is a nonlinear classifier, trained to distinguish order and disorder based on a relatively large feature space (including average coordination number, amino acid compositions (aromatic and charged residues), and net charge). Thus, CH feature space can be considered as a subset of PONDR VLXT feature space [30]. Importantly, these findings may be physically interpretable in terms of different types of disorder, collapsed (molten globule-like) and extended (premolten globule- and coil-like). Under this consideration, the CH-plot classification discriminates proteins with the extended disorder from a set of globular conformations (molten globule-like or rigid well-structured proteins) and proteins predicted to be disordered by the CH-plot approach are likely to belong to the extended disorder class. On the other hand, PONDR-based approaches can discriminate all disordered conformations (coil-like, premolten globules, and molten globules) from rigid well-folded proteins, suggesting that CH classification is roughly a subset of PONDR VL-XT, in both predictions of disorder and feature space [30]. Based on this reasoning, several interesting conclusion have been made. It has been suggested that if a protein is predicted to be disordered by both CH and CDF, then, it is likely to be in the extended disorder class. However, a protein predicted to be disordered by CDF but predicted to be ordered by CH-plot might have properties consistent with a dynamic, collapsed chain; that is, it is likely to be in the native molten globule class. Finally, proteins predicted to be ordered by both algorithms are of course likely to be in the well-structured class [30].

2.2. Estimation of Commonness of Disorder in Various Proteomes. The first application of the disorder predictors was the evaluation of the commonness of protein disorder in the Swiss-Prot database [61]. This analysis revealed that 25% of proteins in Swiss-Prot had predicted ID regions longer than 40 consecutive residues and that at least 11% of residues in Swiss-Prot were likely to be disordered. Given the existence of a few dozen experimentally characterized disordered regions at the time, this work had significant influence on the recognition of the importance of studying disordered proteins [61].

Next, both PONDR VLXT [28] and 3 Flavor PONDR predictors [73] were used to estimate the amount of disorder in various genomes. The predictions counted only disordered regions greater than forty residues in length, which has a false-positive rate of fewer than 400 residues out of 100 000. The result clearly showed that disorder increases from bacteria to archaea to eukaryota with over half of the eukaryotic proteins containing predicted disordered regions [28, 73]. One explanation for this trend is a change in the cellular requirements for certain protein functions, particularly cellular signaling. In support of this hypothesis, PONDR analysis of a eukaryotic signal protein database indicates that the majority of known signal transduction proteins are predicted to contain significant regions of disorder [25].

Ward et al. [29] have refined and systematized such an analysis and concluded that the fraction of proteins containing disordered regions of 30 residues or longer (predicted using DISOPRED) were 2% in archaea, 4% in bacteria, and 33% in eukarya. In addition, a complete functional analysis of the yeast proteome with respect to the three Gene Ontology (GO) categories was performed. In terms of molecular function, transcription, kinase, nucleic acid, and protein-binding activity were the most distinctive signatures of disordered proteins. The most overrepresented GO terms characteristic for the biological process category were transposition, development, morphogenesis, protein phosphorylation, regulation, transcription, and signal transduction. Finally, with respect to cellular component, it appeared that nuclear proteins were significantly enriched in disorder, while terms such as membrane, cytosol, mitochondrion, and cytoplasm were distinctively overrepresented in ordered proteins [29].

Application of the CH-CDF analysis to various proteomes revealed that CDF analysis predicts about 2-fold higher frequency of disorder in sequence databases than CH-plot classification suggesting that approximately half of disordered proteins in different proteomes possess extended disorder, whereas another half represents proteins with the collapsed disorder [30]. Furthermore, the consensus CDF-CH method showed that approximately 4.5% of *Yersinia pestis*, 5% of *Escherichia coli* K12, 6% of *Archaeoglobus fulgidus*, 8% of *Methanobacterium thermoautotrophicum*, 23% of *Arabidopsis thaliana*, and 28% of *Mus musculus* proteins are wholly disordered [30].

As mentioned above, the CH-plot, being a linear classifier, takes into account only two parameters of the particular sequence—charge and hydropathy, whereas CDF analysis is dependent upon the output of the PONDR VLXT predictor, a nonlinear neural network classifier, which was trained to distinguish order and disorder based on a significantly larger feature space that explicitly includes net charge and hydropathy [30]. According to these methodological differences, CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and premolten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). On the other hand, PONDR-based CDF analysis may discriminate all disordered conformations including molten globules from rigid well-folded proteins [30].

This difference in the sensitivity of predictors to different levels of overall disorderedness was utilized in CDF-CH-plot analysis, which allows the ordered and disordered proteins separation in the CH-CDF phase space [92]. In this approach, each spot corresponds to a single protein and its coordinates are calculated as a distance of this protein from the boundary in the corresponding CH-plot (Y-coordinate) and an averaged distance of the corresponding CDF curve from the boundary (X-coordinate). Positive and negative Y values correspond to proteins which, according to CH-plot analysis, are predicted to be natively unfolded or compact, respectively. Whereas positive and negative X values are attributed to proteins that, by the CDF analysis, are predicted to be ordered or intrinsically disordered, respectively. Therefore, this plot has four quadrants: (−,−) quadrant, which contains proteins predicted to be disordered by CDF, but compact by CH-plot (i.e., proteins with molten globule-like properties); (−,+) quadrant that includes proteins predicted

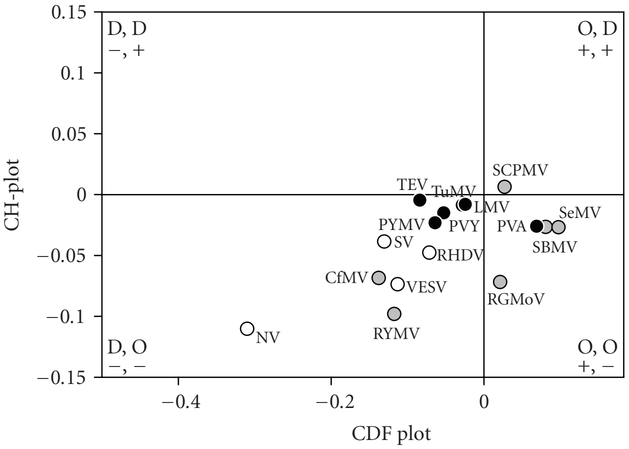


Figure 3: *CH-CDF analysis of the genome-linked proteins VPgs from various viruses.* Comparison of the results of PONDR CDF and CH-plot analyses for whole protein order-disorder via distributions of VPgs within the CH-CDF phase space. The protein analyzed by this approach are VPgs from *Sobemovirus* (Rice yellow mottle virus (RYMV), Cocksfoot mottle virus (CoMV), Ryegrass mottle virus (RGMoV), Southern bean mosaic virus (SBMV), Southern cowpea mosaic virus (SCPMV), Sesbania mottle virus (SeMV)), Potyvirus (Lettuce mosaic virus (LMV), Potato virus Y (PVY), Potato virus A (PVA), Tobacco etch virus (TEV), Turnip mosaic virus (TuMV), Bean yellow mosaic virus (BYMV)), and *Caliciviridae* (Rabbit hemorrhabic disease virus (RHDV), Vesicular exanthema of swine virus (VESV), Man Sapporo virus Manchester virus (SV), and Norwalk virus (NV)).

to be disordered by both methods (i.e., random coils and premolten globules); (+,−) quadrant which contains ordered proteins; and (+,+) quadrant including proteins which are predicted to be disordered by CH-plot, but ordered by the CDF analysis [92]. Application of such a combined CH-CDF analysis to mice proteins revealed that ∼12% mice proteins are likely to belong to the class of extended IDPs (native coils and native premolten globules), whereas ∼30% proteins in mouse genome are potential native molten globules.

Recently, the disorderedness of genome-linked proteins VPg from various viruses were evaluated by a combined CH-CDF analysis [93, 94]. The genome-linked protein VPg of Potato virus A (PVA; genus Potyvirus) has essential functions in all critical steps of PVA infection, that is, replication, movement, and virulence. The structural analysis of the recombinant PVA VPg revealed this protein possesses many properties of a native molten globule [93]. In a follow-up study, it has been shown that although VPgs from various viruses (from *Sobemovirus, Potyvirus*, and *Caliciviridae* genera) are highly diverse in size, sequence, and function, many of them were predicted to contain long disordered domains [94]. Figure 3 summarizes these findings by showing the localization of several VPgs within the CH-CDF phase space. Each spot represents a single VPg whose coordinates were calculated as a distance of this protein from the boundary in the corresponding charge-hydropathy plot (CH-plot, Y-coordinate) and an averaged distance of the corresponding cumulative distribution function (CDF) curve from the boundary (X-coordinate). Figure 3 clearly shows that many VPgs are expected to behave as native molten globules [94].

Several recent reviews summarized the current state of the art in the field of IDP predictions and represent a useful overview of the prediction methods highlighting their advantages and drawbacks [46, 91]. Concluding, the considered-above studies clearly showed that IDPs (both collapsed and extended) are highly abundant in nature and can be reliably identified by various computational means.

2.3. Discovering the Disease-Related Unfoldomes.

Misfolding (the failure of a specific peptide or protein to adopt its functional conformational state) and related dysfunction of many proteins were considered as a major cause for the development of different pathological conditions. Such misfolding and dysfunction can originate from point mutation(s) or result from an exposure to internal or external toxins, impaired posttranslational modifications (phosphorylation, advanced glycation, deamidation, racemization, etc.), an increased probability of degradation, impaired trafficking, lost binding partners, or oxidative damage. All these factors can act independently or in association with one another.

Although the formation of various aggregates represents the most visible consequence of protein misfolding and although these aggregates form the basis for the development of various protein deposition diseases, pathogenesis of many more human diseases does not depend on aggregation being based on protein dysfunction. As many of the proteins associated with the conformational diseases are also involved in recognition, regulation, and cell signaling, it has been hypothesized that many of them are IDPs. In other words, according to this the "disorder in disorders" or D^2 concept, IDPs are abundantly involved in the development of the conformational diseases, which therefore may originate from the misidentification, misregulation, and missignaling due to the misfolding of causative IDPs [39].

To support this hypothesis, three approaches were elaborated for estimating the abundance of IDPs in various pathological conditions. The first approach is based on the assembly of specific datasets of proteins associated with a given disease and the computational analysis of these datasets using a number of disorder predictors [25, 39, 95, 96]. In essence, this is an analysis of individual proteins extended to a set of independent proteins. A second approach utilized network of genetic diseases where the related proteins are interlinked within one disease and between different diseases [41]. A third approach is based on the evaluation of the association between a particular protein function (including the disease-specific functional keywords) with the level of intrinsic disorder in a set of proteins known to carry out this function [13–15]. These three approaches are briefly presented below.

The easiest way to evaluate the abundance of intrinsic disorder in a given disease is based on a simple two-stage protocol, where a set of disease-related proteins is first assembled by searching various databases and then the collected group of proteins is analyzed for intrinsic disorder. The depth of this analysis is based on the breadth of the search for the disease-related proteins and on the number of different computational tools utilized to find disordered proteins/regions [25, 39, 44, 95–97]. Using this approach, it has been shown that many proteins associated with cancer, neurodegenerative diseases, and cardiovascular disease are highly disordered, being depleted in major order-promoting residues (Trp, Phe, Tyr, Ile, and Val) and enriched in some disorder-promoting residues (Arg, Gln, Ser, Pro, and Glu). High level of intrinsic disorder and a substantial number of potential interaction sites were also found using a set of computational tools. Many proteins were predicted to be wholly disordered. Overall, these studies clearly showed that intrinsic disorder is highly prevalent in proteins associated with human diseases, being comparable with that of signaling proteins and significantly exceeding the levels of intrinsic disorder in eukaryotic and in nonhomologous, structured proteins.

Unfoldome of human genetic diseases was assembled via the analysis of a specific network which was built to estimate whether human genetic diseases and the corresponding disease genes are related to each other at a higher level of cellular and organism organization. This network represented a bipartite graph with a network of genetic diseases, the human disease network (HDN), where two diseases were directly linked if there was a gene that was directly related to both of them, and a network of disease genes, the disease gene network (DGN), where two genes were directly linked if there was a disease to which they were both directly related [98]. This framework, called the human diseasome, systematically linked the human disease phenome (which includes all the human genetic diseases) with the human disease genome (which contains all the disease-related genes) [98]. The analysis of HDN revealed that of 1284 genetic diseases, 867 had at least one link to other diseases, and 516 diseases formed a giant component, suggesting that the genetic origins of most diseases, to some extent, were shared with other diseases. In the DGN, 1377 of 1777 disease genes were shown to be connected to other disease genes, and 903 genes belonged to a giant cluster HDN. The vast majority of genes associated with genetic diseases was nonessential and showed no tendency to encode hub proteins (i.e., proteins having multiple interactions) [98]. The large-scale analysis of the abundance of intrinsic disorder in transcripts of the various disease-related genes was performed using a set of computational tools which uncovers several important features [41, 44]: (a) intrinsic disorder is common in proteins associated with many human genetic diseases; (b) different disease classes vary in the IDP contents of their associated proteins; (c) molecular recognition features, which are relatively short loosely structured protein regions within mostly disordered sequences and which gain structure upon binding to partners, are common in the diseasome, and their abundance correlates with the intrinsic disorder level; (d) some disease classes have a significant fraction of genes affected by alternative splicing, and the alternatively spliced regions in the corresponding proteins are predicted to be highly disordered and in some diseases contain a significant number of molecular recognition features, MoRFs; (e) correlations were found among the various diseasome graph-related properties and intrinsic disorder. In agreement with earlier studies, hub proteins were shown to be more disordered.

Another approach is a computational tool elaborated for the evaluation of a correlation between the functional annotations in the SWISSPROT database and the predicted intrinsic disorder was elaborated [13–15]. The approach is based on the hypothesis that if a function described by a given keyword relies on intrinsic disorder, then the keyword-associated protein would be expected to have a greater level of predicted disorder compared to the protein randomly chosen from the SWISSPROT. To test this hypothesis, functional keywords associated with 20 or more proteins in SWISSPROT were found and corresponding keyword-associated datasets of proteins were assembled. Next, for each such a keyword-associated set, a length-matching set of random proteins was drawn from the SWISSPROT, and order-disorder predictions were carried out for the keyword-associated sets and for the random sets [13–15]. The application of this tool revealed that out of 710 SWISSPROT keywords, 310 functional keywords were associated with ordered proteins, 238 functional keywords were attributed to disordered proteins, and the remainder 162 keywords yield ambiguity in the likely function-structure associations [13–15]. It has been also shown that keywords describing various diseases were strongly correlated with proteins predicted to be disordered. Contrary to this, no disease-associated proteins were found to be strongly correlated with absence of disorder [14].

3. Experimental Tools for the Unfoldome Discovery

3.1. Enrichment of Cell Extracts in Extended IPDs by Acid Treatment. Extended IDPs, being characterized by high percentages of charged residues, do not undergo large-scale structural changes at low pH [55]. As a result, many of these proteins were shown to remain soluble under these extreme conditions [99, 100]. On the contrary, the protonation of negatively charged side chains in ordered proteins is commonly accompanied by protein denaturation or unfolding [101–104]. Unlike IDPs, the acidic pH-induced denatured conformations of structured proteins contain larger number of hydrophobic residues. The pH-induced exposure of these normally buried hydrophobic residues makes "A" states (i.e., partially folded conformations induced by acidic pH) of globular proteins "sticky," leading to their aggregation and precipitation.

Analysis or literature data revealed a set of 29 proteins which do not precipitate during perchloric acid (PCA) or trichloroacetic acid (TCA) treatment of cell extracts and this resistance to PCA or TCA treatment was utilized for their isolation [40]. However, 14 of these PCA/TCA-soluble proteins were experimentally determined to be totally unstructured, 6 were structured, and 9 had not been structurally characterized, suggesting that at least 50% of the proteins isolated by virtue of their resistance to PCA or TCA could be expected to be totally unstructured [40]. To gain more information on the abundance of intrinsic disorder in acid-soluble proteins, their sequences were analyzed using two binary predictors of intrinsic disorder, CH-plot [45] and CDF analysis [30], both of which perform binary classification of whole proteins as either mostly disordered or mostly ordered, where mostly ordered indicates proteins that contain more ordered residues than disordered residues and mostly disordered indicates proteins that contain more disordered residues than ordered residues. The results of this analysis revealed an excellent correlation between experiment and prediction: the majority of proteins experimentally shown to be structured or unfolded were predicted to be ordered or intrinsically disordered, respectively, by both predictors. Additionally, three of four experimentally uncharacterized proteins were predicted to be wholly disordered by both classifiers. Thus, a combination of experimental and computational approaches suggested that ∼70% of acid soluble proteins isolated based on their resistance to PCA or TCA could be expected to be totally unstructured [40].

Based on these observations it was suggested that indifference to acid treatment represents one of the characteristic properties of extended IDPs, which can result in the substantial enrichment of IDPs in the soluble fraction after the acid treatment, and, therefore, can be exploited to develop standard protocols for isolating and studying IDPs on a proteomic scale [40]. In agreement with this hypothesis, treatment of *E.coli* cell extracts with 1% PCA resulted in a total protein reduction of ∼30 000-fold when compared to the total soluble extract, and 3% PCA was sufficient to denature and precipitate all nonresistant proteins because higher PCA concentrations did not result in further yield reductions [40]. Treatment with 3% TCA resulted in a yield similar to 1% PCA. The acid-soluble fractions from the *E.coli* extracts were visualized using 2D SDS-PAGE, which revealed that a substantial number of *E. coli* proteins were resistant to acid denaturation and concomitant precipitation. In fact, this analysis revealed that 158 proteins remained soluble in the presence of 5% PCA [40]. This suggests that ∼110 of the PCA-soluble proteins could be expected to be totally unstructured (based on the assumption that ∼ 70% acid-stable proteins are totally unstructured, see above). This number compares favorably with the 85 to 196 totally disordered proteins estimated to be present in the *E. coli* proteome [30]. Therefore, treating total protein extracts with 3–5% PCA or TCA and determining the identities of the soluble proteins could form the basis for uncovering unfoldomes in various organisms.

3.2. Enrichment of Cell Extracts in Extended IPDs by Heat Treatment. Several IDPs were shown to possess high resistance toward heat denaturation and aggregation. In fact, the solubility and limited secondary structure of extended IDPs, such as p21, p27, α-synuclein, prothymosin α, and phosphodiesterase γ subunit, were virtually unaltered by heating to 90°C [31, 32, 63, 100, 105–111]. This resistance to thermal aggregation, which likely originates from the low mean hydrophobicity and high net charge characteristic of extended IDPs, has been utilized for the purification of these proteins [63, 112–115]. The indifference to heat treatment was proposed as an analytical tool for the evaluation of the abundance of extended IDPs in various proteomes [30, 57].

Recently, the extracts of NIH3T3 mouse fibroblasts were heated at a variety of temperatures and analyzed by SDS−PAGE to determine the extent of protein precipitation under these conditions [57]. In agreement with previous studies, this analysis revealed that the increase in the incubation temperature was accompanied by the decrease in the amount of soluble proteins. In fact, 375, 388, and 198 proteins (287, 304, and 124 nonredundant proteins) were identified by MALDI-TOF/TOF mass spectrometry from 584, 472, and 269 spots on 2D gels obtained for cell extracts treated at 4, 60, and 98°C, respectively. These nonredundant proteins were further analyzed using a set of bioinformatics tools. In this study, proteins were classified as IDPs (proteins having an average PONDR VLXT score >0.5 and proteins having an average PONDR score of 0.32–0.5 and possessing a high-mean net charge and low-mean hydrophobicity), intrinsically folded proteins (IFPs, proteins having an average PONDR score <0.32), or mixed ordered/disordered proteins (MPs, proteins that did not meet the described above criteria for IUPs or IFPs) [57]. The analysis clearly showed that heat treatment resulted in an enrichment of IDPs and depletion of MPs and IFPs. In fact, although IDPs comprised only 11.8% of the proteins identified in the untreated cell extract (4°C), their relative population increased to 41.9% after the heat treatment at 98°C. On the other hand, MPs and IFPs, which comprised 42.8 and 45.4% of proteins in the untreated cell extract, were substantially depleted to 27.4 and 30.6%, respectively, after heat treatment at 98°C [57].

3.3. Finding IDPs by the Combination of Native and 8 M Urea Electrophoresis of Heat-Treated Proteins. The fact that extended IDPs are characterized by heat stability and structural indifference to chemical denaturation was recently utilized in novel 2-D gel-electrophoresis technique which consists of the combination of native and 8 M urea electrophoresis of heat-treated proteins [56]. The rationales for this approach are considered below. As discussed above, extended IDPs are often heat-stable as demonstrated for Csd1 [112], MAP2 [114], α-synuclein [63, 107], its familial Parkinson's disease-related mutants [108], β- and γ-synucleins [110], stathmin [113], p21$^{\mathrm{Cip\,1}}$ [31], prothymosin α [100], C-terminal domain of caldesmon [111], and phosphodiesterase γ [109]. Therefore, heat treatment should lead to a decent initial separation of the extended IDPs from globular proteins, the vast majority of which are known to aggregate and precipitate at high temperatures. In the native gel, IDPs and rare heat-stable globular proteins will then be separated according to their charge/mass ratios. Since the extended IDPs are as unfolded in 8 M urea as under native conditions, they are expected to run the same distance in the second dimension and end up along the diagonal. Heat-stable globular proteins will unfold in urea, slow down in the second direction due to the increased size, and therefore will accumulate above the diagonal (see Figure 4 for the schematic representation of this technique). This difference

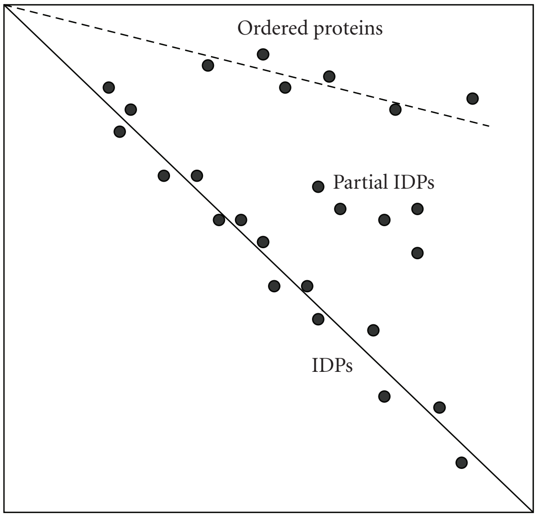


Figure 4: Schematic representation of the native/8 M urea 2D electrophoresis for separation of extended IDPs and globular proteins. A *continuous line* marks the diagonal of the gel to where IDPs run. A *dashed line* marks the position of globular proteins. A few proteins with a mixture of ordered and disordered regions are also indicated as “Partial IDPs”.

in conformational behavior between ordered proteins and IDPs can lead to their effective separation, enabling the IDP identification by mass-spectrometry [56].

The usefulness of this approach has been validated via the analysis of a set of 10 experimentally characterized IDPs (stathmin, MAP2c, Mypt1-(304–511), ERD10, α-casein, β-casein, α-synuclein, CSD1, Bob-1, and DARPP32) and a set of 4 globular control proteins (fetuin, IPMDH, BSA, ovalbumin) [56]. The analysis revealed that IDPs ran at, or very near, the diagonal of the second (denaturing) gel, whereas globular proteins remained way above the diagonal, clearly showing that the proposed 2D electrophoresis is able to separate IDPs and globular proteins as predicted [56].

Next, heat-treated extracts of *E. coli* and *S. cerevisiae* were analyzed by this 2D electrophoresis [56]. This analysis revealed that more *S. cerevisiae* proteins were seen in the diagonal in agreement with predictions that the frequency of protein disorder increases with increasing complexity of the organisms [5, 28–30]. At the next step, some spots at and above the diagonal were identified by mass-spectrometry and the intrinsic disorder propensity of the identified proteins was estimated by PONDR VLXT [56]. This analysis revealed that the amount of predicted disorder in proteins located at the diagonal positions was very high (52.1 ± 14.1%), noticeably exceeding that of typical IDPs such as α-synuclein (37.1%) and α-casein (41.15%) [56]. Although many of the “diagonal” proteins have never been structurally characterized, literature data were available for some of them. The list of such previously characterized IDPs identified in this study includes ribosomal proteins, GroES, and acyl carrier protein. The majority of proteins above the diagonal were found to be enzymes (e.g., superoxide dismutase), which are known to require a well-defined structure for function [56]. Based on these finding it has been concluded that the proposed 2D electrophoresis is suitable for the proteome-wide identification of IDPs.

4. Concluding Remarks

Intrinsic disorder is highly abundant in nature. According to the genome-based bioinformatics predictions, significant fraction of any given proteome belongs to the class of IDPs. These proteins possess numerous vital functions. Many proteins associated with various human diseases are intrinsically disordered too. High degree of association between protein intrinsic disorder and maladies is due to structural and functional peculiarities of IDPs and IDRs, which are typically involved in cellular regulation, recognition, and signal transduction. As the number of IDPs is very large and as many of these proteins are interlinked, the concepts of the unfoldome and unfoldomics were introduced. IDPs, especially their extended forms, are characterized by several unique features that can be used for isolation of these proteins from the cell extracts. The corresponding proteomic techniques utilize specific high resistance of IDPs against extreme pH and high temperature, as well as their structural indifference to chemical denaturation. At the computational side, several specific features of the IDP amino acid sequences provide a solid background for the reliable identification of these proteins at the proteome level. These proteomic-scale identification and characterization of IDPs are needed to advance our knowledge in this important field.

Acknowledgments

This review would not be possible without many colleagues who, over the years, contributed to the introduction and development of the IDP field. Numerous very helpful and thought-promoting discussions with my colleagues and fiends are deeply acknowledged. This work was supported in part by the Program of the Russian Academy of Sciences for the “Molecular and cellular biology,” and by grants R01 LM007688-01A1 and GM071714-01A2 from the National Institutes of Health. The author gratefully acknowledges the support of the IUPUI Signature Centers Initiative.

References

[1] E. Fischer, “Einfluss der configuration auf die wirkung der enzyme,” *Berichte der Deutschen Chemischen Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, 1894.

[2] R. U. Lemieux and U. Spohr, “How Emil Fischer was led to the lock and key concept for enzyme specificity,” *Advances in Carbohydrate Chemistry and Biochemistry*, vol. 50, pp. 1–20, 1994.

[3] C. M. Dobson, “Protein misfolding, evolution and disease,” *Trends in Biochemical Sciences*, vol. 24, no. 9, pp. 329–332, 1999.

[4] A. K. Dunker, C. J. Oldfield, J. Meng, et al., “The unfoldomics decade: an update on intrinsically disordered proteins,” *BMC Genomics*, vol. 9, supplement 2, article S1, 2008.

[5] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[6] A. K. Dunker, C. J. Brown, and Z. Obradović, "Identification and functions of usefully disordered proteins," *Advances in Protein Chemistry*, vol. 62, pp. 25–49, 2002.

[7] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky, "Flexible nets: the roles of intrinsic disorder in protein interaction networks," *FEBS Journal*, vol. 272, no. 20, pp. 5129–5148, 2005.

[8] A. K. Dunker, J. D. Lawson, C. J. Brown, et al., "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26–59, 2001.

[9] A. K. Dunker and Z. Obradović, "The protein trinity—linking function and disorder," *Nature Biotechnology*, vol. 19, no. 9, pp. 805–806, 2001.

[10] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradović, V. N. Uversky, and A. K. Dunker, "Intrinsic disorder and functional proteomics," *Biophysical Journal*, vol. 92, no. 5, pp. 1439–1456, 2007.

[11] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Current Opinion in Structural Biology*, vol. 18, no. 6, pp. 756–764, 2008.

[12] A. K. Dunker and V. N. Uversky, "Signal transduction via unstructured protein conduits," *Nature Chemical Biology*, vol. 4, no. 4, pp. 229–230, 2008.

[13] S. Vucetic, H. Xie, L. M. Iakoucheva, et al., "Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1899–1916, 2007.

[14] H. Xie, S. Vucetic, L. M. Iakoucheva, et al., "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1917–1932, 2007.

[15] H. Xie, S. Vucetic, L. M. Iakoucheva, et al., "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1882–1898, 2007.

[16] M. S. Cortese, V. N. Uversky, and A. Keith Dunker, "Intrinsic disorder in scaffold proteins: getting more from less," *Progress in Biophysics and Molecular Biology*, vol. 98, no. 1, pp. 85–106, 2008.

[17] R. B. Russell and T. J. Gibson, "A careful disorderliness in the proteome: sites for interaction and targets for future therapies," *FEBS Letters*, vol. 582, no. 8, pp. 1271–1275, 2008.

[18] C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Qu, V. N. Uversky, and A. K. Dunker, "Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners," *BMC Genomics*, vol. 9, supplement 1, article S1, 2008.

[19] P. Tompa and P. Csermely, "The role of structural disorder in the function of RNA and protein chaperones," *FASEB Journal*, vol. 18, no. 11, pp. 1169–1175, 2004.

[20] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling," *Journal of Molecular Recognition*, vol. 18, no. 5, pp. 343–384, 2005.

[21] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 321–331, 1999.

[22] P. Tompa, "Intrinsically unstructured proteins," *Trends in Biochemical Sciences*, vol. 27, no. 10, pp. 527–533, 2002.

[23] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Science*, vol. 11, no. 4, pp. 739–756, 2002.

[24] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 3, pp. 197–208, 2005.

[25] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins," *Journal of Molecular Biology*, vol. 323, no. 3, pp. 573–584, 2002.

[26] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Current Opinion in Structural Biology*, vol. 18, no. 6, pp. 756–764, 2008.

[27] P. Tompa, "The interplay between structure and function in intrinsically unstructured proteins," *FEBS Letters*, vol. 579, no. 15, pp. 3346–3354, 2005.

[28] A. K. Dunker, Z. Obradović, P. Romero, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Informatics*, vol. 11, pp. 161–171, 2000.

[29] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[30] C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Bunker, "Comparing and combining predictors of mostly disordered proteins," *Biochemistry*, vol. 44, no. 6, pp. 1989–2000, 2005.

[31] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright, "Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 21, pp. 11504–11509, 1996.

[32] E. R. Lacy, I. Filippov, W. S. Lewis, et al., "p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding," *Nature Structural and Molecular Biology*, vol. 11, no. 4, pp. 358–364, 2004.

[33] E. R. Lacy, Y. Wang, J. Post, et al., "Molecular basis for the specificity of p27 toward cyclin-dependent kinases that regulate cell division," *Journal of Molecular Biology*, vol. 349, no. 4, pp. 764–773, 2005.

[34] H. J. Dyson and P. E. Wright, "Coupling of folding and binding for unstructured proteins," *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 54–60, 2002.

[35] C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky, and A. K. Dunker, "Coupled folding and binding with alpha-helix-forming molecular recognition elements," *Biochemistry*. In press.

[36] Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky, and A. K. Dunker, "Mining α-helix-forming molecular recognition features with cross species sequence alignments," *Biochemistry*, vol. 46, no. 47, pp. 13468–13477, 2007.

[37] A. Mohan, *MoRFs: A Dataset of Molecular Recognition Features*, The School of Informatics, Indiana University, Indianapolis, Ind, USA, 2006.

[38] V. Vacic, C. J. Oldfield, A. Mohan, et al., "Characterization of molecular recognition features, MoRFs, and their binding partners," *Journal of Proteome Research*, vol. 6, no. 6, pp. 2351–2366, 2007.

[39] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: introducing the D2 concept," *Annual Review of Biophysics*, vol. 37, pp. 215–246, 2008.

[40] M. S. Cortese, J. P. Baird, V. N. Uversky, and A. K. Dunker, "Uncovering the unfoldome: enriching cell extracts for unstructured proteins by acid treatment," *Journal of Proteome Research*, vol. 4, no. 5, pp. 1610–1618, 2005.

[41] U. Midic, C. J. Oldfield, A. K. Keith, Z. Obradović, and V. N. Uversky, "Protein disorder in the human diseasome: unfoldomics of human genetic diseases," *BMC Genomics*, vol. 10, supplement 1, article S12, 2009.

[42] H. Winkler, V.u.U.d.P.i.P.-u.T.J.V.F., 1920.

[43] J. Lederberg and A. T. McCray, "'Ome' sweet 'omics'—a genealogical treasury of words," *The Scientist*, vol. 15, no. 7, article 8, 2001.

[44] V. N. Uversky, C. J. Oldfield, U. Midic, et al., "Unfoldomics of human diseases: linking protein intrinsic disorder with diseases," *BMC Genomics*, vol. 10, supplement 1, article S7, 2009.

[45] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why are "natively unfolded" proteins unstructured under physiologic conditions?" *Proteins*, vol. 41, no. 3, pp. 415–427, 2000.

[46] F. Ferron, S. Longhi, B. Canard, and D. Karlin, "A practical overview of protein disorder prediction methods," *Proteins*, vol. 65, no. 1, pp. 1–14, 2006.

[47] Z. Dosztanyi, M. Sandor, P. Tompa, and I. Simon, "Prediction of protein disorder at the domain level," *Current Protein and Peptide Science*, vol. 8, no. 2, pp. 161–171, 2007.

[48] Z. Dosztanyi and P. Tompa, "Prediction of protein disorder," *Methods in Molecular Biology*, vol. 426, pp. 103–115, 2008.

[49] S. J. Hubbard, R. J. Beynon, and J. M. Thornton, "Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures," *Protein Engineering*, vol. 11, no. 5, pp. 349–359, 1998.

[50] L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, R. D. Smith, A. K. Dunker, and E. J. Ackerman, "Aberrant mobility phenomena of the DNA repair protein XPA," *Protein Science*, vol. 10, no. 7, pp. 1353–1362, 2001.

[51] R. Reeves and M. S. Nissen, "Purification and assays for high mobility group HMG-I(Y) protein function," *Methods in Enzymology*, vol. 304, pp. 155–188, 1999.

[52] G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker, "Natively disordered proteins," in *Handbook of Protein Folding*, J. Buchner and T. Kiefhaber, Eds., pp. 271–353, Wiley-VCH, Weinheim, Germany, 2005.

[53] V. Receveur-Bréchot, J.-M. Bourhis, V. N. Uversky, B. Canard, and S. Longhi, "Assessing protein disorder and induced folding," *Proteins*, vol. 62, no. 1, pp. 24–45, 2006.

[54] V. N. Uversky, "What does it mean to be natively unfolded?" *European Journal of Biochemistry*, vol. 269, no. 1, pp. 2–12, 2002.

[55] V. N. Uversky, "Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, ph, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding," *Protein Journal*, vol. 28, no. 7-8, pp. 305–325, 2009.

[56] V. Csizmók, E. Szollosi, P. Friedrich, and P. Tompa, "A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins," *Molecular and Cellular Proteomics*, vol. 5, no. 2, pp. 265–273, 2006.

[57] C. A. Galea, V. R. Pagala, J. C. Obenauer, C.-G. Park, C. A. Slaughter, and R. W. Kriwacki, "Proteomic studies of the intrinsically unstructured mammalian proteome," *Journal of Proteome Research*, vol. 5, no. 10, pp. 2839–2848, 2006.

[58] P. Romero, Z. Obradović, C. Kissinger, J. E. Villafranca, and A. K. Dunker, "Identifying disordered regions in proteins from amino acid sequence," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 1, pp. 90–95, 1997.

[59] P. Romero, Z. Obradović, and K. Dunker, "Sequence data analysis for long disordered regions prediction in the calcineurin family," *Genome Informatics*, vol. 8, pp. 110–124, 1997.

[60] Q. Xie, G. E. Arnold, P. Romero, Z. Obradović, E. Garner, and A. K. Dunker, "The sequence attribute method for determining relationships between sequence and protein disorder," *Genome Informatics*, vol. 9, pp. 193–200, 1998.

[61] P. Romero, Z. Obradović, C. R. Kissinger, et al., "Thousands of proteins likely to have long disordered regions," *Pacific Symposium on Biocomputing*, pp. 437–448, 1998.

[62] K. Gast, H. Damaschun, K. Eckert, et al., "Prothymosin α: a biologically active protein with random coil conformation," *Biochemistry*, vol. 34, no. 40, pp. 13211–13218, 1995.

[63] P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, and P. T. Lansbury Jr., "NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded," *Biochemistry*, vol. 35, no. 43, pp. 13709–13715, 1996.

[64] H. C. Hemmings Jr., A. C. Nairn, D. W. Aswad, and P. Greengard, "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," *Journal of Neuroscience*, vol. 4, no. 1, pp. 99–110, 1984.

[65] R. J. Williams, "The conformational mobility of proteins and its functional significance," *Biochemical Society Transactions*, vol. 6, no. 6, pp. 1123–1126, 1978.

[66] A. K. Dunker, E. Garner, S. Guilliot, et al., "Protein disorder and the evolution of molecular recognition: theory, predictions and observations," *Pacific Symposium on Biocomputing*, pp. 473–484, 1998.

[67] E. Garner, P. Cannon, P. Romero, Z. Obradović, and A. K. Dunker, "Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization," *Genome Informatics*, vol. 9, pp. 201–213, 1998.

[68] M. Vihinen, E. Torkkila, and P. Riikonen, "Accuracy of protein flexibility predictions," *Proteins*, vol. 19, no. 2, pp. 141–149, 1994.

[69] R. M. Williams, Z. Obradović, V. Mathura, et al., "The protein non-folding problem: amino acid determinants of intrinsic order and disorder," *Pacific Symposium on Biocomputing*, pp. 89–100, 2001.

[70] P. Romero, Z. Obradović, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins*, vol. 42, no. 1, pp. 38–48, 2001.

[71] X. Li, Z. Obradović, C. J. Brown, E. C. Garner, and A. K. Dunker, "Comparing predictors of disordered protein," *Genome Informatics*, vol. 11, pp. 172–184, 2000.

[72] P. Radivojac, Z. Obradović, D. K. Smith, et al., "Protein flexibility and intrinsic disorder," *Protein Science*, vol. 13, no. 1, pp. 71–80, 2004.

[73] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradović, "Flavors of protein disorder," *Proteins*, vol. 52, no. 4, pp. 573–584, 2003.

[74] X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradović, "Predicting protein disorder for N-, C-, and internal regions," *Genome Informatics*, vol. 10, pp. 30–40, 1999.

[75] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradović, "Optimizing long intrinsic disorder predictors with protein evolutionary information," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 1, pp. 35–60, 2005.

[76] Z. Obradović, K. Peng, S. Vucetic, P. Radivojac, and A. K. Bunker, "Exploiting heterogeneous sequence properties improves prediction of protein disorder," *Proteins*, vol. 61, supplement 7, pp. 176–182, 2005.

[77] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradović, "Length-dependent prediction of protein in intrinsic disorder," *BMC Bioinformatics*, vol. 7, article 208, 2006.

[78] J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, et al., "FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded," *Bioinformatics*, vol. 21, no. 16, pp. 3435–3438, 2005.

[79] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "GlobPlot: exploring protein sequences for globularity and disorder," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3701–3708, 2003.

[80] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure*, vol. 11, no. 11, pp. 1453–1459, 2003.

[81] D. T. Jones and J. J. Ward, "Prediction of disordered regions in proteins from position specific score matrices," *Proteins*, vol. 53, supplement 6, pp. 573–578, 2003.

[82] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder," *Bioinformatics*, vol. 20, no. 13, pp. 2138–2139, 2004.

[83] K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones, "Protein structure prediction servers at University College London," *Nucleic Acids Research*, vol. 33, supplement 2, pp. W36–W38, 2005.

[84] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," *Bioinformatics*, vol. 21, no. 16, pp. 3433–3434, 2005.

[85] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov, "FoldUnfold: web server for the prediction of disordered regions in protein chain," *Bioinformatics*, vol. 22, no. 23, pp. 2948–2949, 2006.

[86] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf, "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins," *Bioinformatics*, vol. 21, no. 16, pp. 3369–3376, 2005.

[87] C.-T. Su, C.-Y. Chen, and Y.-Y. Ou, "Protein disorder prediction by condensed PSSM considering propensity for order or disorder," *BMC Bioinformatics*, vol. 7, article 319, 2006.

[88] C. T. Su, C. Y. Chen, and C. M. Hsu, "iPDA: integrated protein disorder analyzer," *Nucleic Acids Research*, vol. 35, web server issue, pp. W465–W472, 2007.

[89] A. Vullo, O. Bortolamil, G. Pollastri, and S. C. E. Tosatto, "Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines," *Nucleic Acids Research*, vol. 34, web server issue, pp. W164–W168, 2006.

[90] T. Ishida and K. Kinoshita, "PrDOS: prediction of disordered protein regions from amino acid sequence," *Nucleic Acids Research*, vol. 35, web server issue, pp. W460–W464, 2007.

[91] B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, and A. K. Dunker, "Predicting intrinsic disorder in proteins: an overview," *Cell Research*, vol. 19, no. 8, pp. 929–949, 2009.

[92] A. Mohan, W. J. Sullivan Jr., P. Radivojac, A. K. Dunker, and V. N. Uversky, "Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes," *Molecular BioSystems*, vol. 4, no. 4, pp. 328–340, 2008.

[93] K. I. Rantalainen, V. N. Uversky, P. Permi, N. Kalkkinen, A. K. Dunker, and K. Mäkinen, "Potato virus A genome-linked protein VPg is an intrinsically disordered molten globule-like protein with a hydrophobic core," *Virology*, vol. 377, no. 2, pp. 280–288, 2008.

[94] E. Hébrard, Y. Bessin, T. Michon, et al., "Intrinsic disorder in Viral Proteins Genome-Linked: experimental and predictive analyses," *Virology Journal*, vol. 6, article 23, 2009.

[95] V. N. Uversky, A. Roman, C. J. Oldfield, and A. K. Dunker, "Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs," *Journal of Proteome Research*, vol. 5, no. 8, pp. 1829–1842, 2006.

[96] Y. Cheng, T. LeGall, C. J. Oldfield, A. K. Dunker, and V. N. Uversky, "Abundance of intrinsic disorder in protein associated with cardiovascular disease," *Biochemistry*, vol. 45, no. 35, pp. 10448–10460, 2006.

[97] V. N. Uversky, "Intrinsic disorder in proteins associated with neurodegenerative diseases," *Frontiers in Bioscience*, vol. 14, pp. 5188–5238, 2009.

[98] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[99] V. N. Uversky, "A protein-chameleon: conformational plasticity of α-synuclein, a disordered protein involved in neurodegenerative disorders," *Journal of Biomolecular Structure and Dynamics*, vol. 21, no. 2, pp. 211–234, 2003.

[100] V. N. Uversky, J. R. Gillespie, I. S. Millett, et al., "Natively unfolded human prothymosin α adopts partially folded collapsed conformation at acidic pH," *Biochemistry*, vol. 38, no. 45, pp. 15009–15016, 1999.

[101] Y. Goto, L. J. Calciano, and A. L. Fink, "Acid-induced folding of proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 2, pp. 573–577, 1990.

[102] Y. Goto and A. L. Fink, "Phase diagram for acidic conformational states of apomyoglobin," *Journal of Molecular Biology*, vol. 214, no. 4, pp. 803–805, 1990.

[103] Y. Goto, N. Takahashi, and A. L. Fink, "Mechanism of acid-induced folding of proteins," *Biochemistry*, vol. 29, no. 14, pp. 3480–3488, 1990.

[104] A. L. Fink, L. J. Calciano, Y. Goto, T. Kurotsu, and D. R. Palleros, "Classification of acid denaturation of proteins: intermediates and unfolded states," *Biochemistry*, vol. 33, no. 41, pp. 12504–12511, 1994.

[105] E. A. Bienkiewicz, J. N. Adkins, and K. J. Lumb, "Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27Kip1," *Biochemistry*, vol. 41, no. 3, pp. 752–759, 2002.

[106] L. Hengst, V. Dulic, J. M. Slingerland, E. Lees, and S. I. Reed, "A cell cycle-regulated inhibitor of cyclin-dependent kinases,"

Proceedings of the National Academy of Sciences of the United States of America, vol. 91, no. 12, pp. 5291–5295, 1994.

[107] V. N. Uversky, J. Li, and A. L. Fink, "Evidence for a partially folded intermediate in α-synuclein fibril formation," *Journal of Biological Chemistry*, vol. 276, no. 14, pp. 10737–10744, 2001.

[108] J. Li, V. N. Uversky, and A. L. Fink, "Effect of familial Parkinson's disease point mutations A30P and A53T on the structural properties, aggregation, and fibrillation of human α-synuclein," *Biochemistry*, vol. 40, no. 38, pp. 11604–11613, 2001.

[109] V. N. Uversky, S. E. Permyakov, V. E. Zagranichny, et al., "Effect of zinc and temperature on the conformation of the γ subunit of retinal phosphodiesterase: a natively unfolded protein," *Journal of Proteome Research*, vol. 1, no. 2, pp. 149–159, 2002.

[110] V. N. Uversky, J. Li, P. Souillac, et al., "Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of α-synuclein assembly by β- and γ-synucleins," *Journal of Biological Chemistry*, vol. 277, no. 14, pp. 11970–11978, 2002.

[111] S. E. Permyakov, I. S. Millett, S. Doniach, E. A. Permyakov, and V. N. Uversky, "Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin," *Proteins*, vol. 53, no. 4, pp. 855–862, 2003.

[112] M. Häckel, T. Konno, and H.-J. Hinz, "A new alternative method to quantify residual structure in 'unfolded' proteins," *Biochimica et Biophysica Acta*, vol. 1479, no. 1-2, pp. 155–165, 2000.

[113] L. D. Belmont and T. J. Mitchison, "Identification of a protein that interacts with tubulin dimers and increases the catastrophe rate of microtubules," *Cell*, vol. 84, no. 4, pp. 623–631, 1996.

[114] M. A. Hernandez, J. Avila, and J. M. Andreu, "Physicochemical characterization of the heat-stable microtubule-associated protein MAP2," *European Journal of Biochemistry*, vol. 154, no. 1, pp. 41–48, 1986.

[115] C. Kalthoff, "A novel strategy for the purification of recombinantly expressed unstructured protein domains," *Journal of Chromatography B*, vol. 786, no. 1-2, pp. 247–254, 2003.