Expanding the Range of Protein Function at the Far End of the Order-Structure Continuum*

Virginia M. Burger 5, Diego O. Nolasco 5, and Collin M. Stultz 5, 9, 1

From the 5Research Laboratory for Electronics, Department of Electrical Engineering & Computer Science, and the 5Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02138

The traditional view of the structure-function paradigm is that a protein’s function is inextricably linked to a well defined, three-dimensional structure, which is determined by the protein’s primary amino acid sequence. However, it is now accepted that a number of proteins do not adopt a unique tertiary structure in solution and that some degree of disorder is required for many proteins to perform their prescribed functions. In this review, we highlight how a number of protein functions are facilitated by intrinsic disorder and introduce a new protein structure taxonomy that is based on quantifiable metrics of a protein’s disorder.

How Do Folded Proteins Differ from Unfolded Proteins?

Proteins are dynamic biological molecules that are involved in virtually all cellular processes (1). At physiologic temperatures, proteins, like all polymers, sample a range of conformations that are a function of the macromolecular environment and the primary amino acid sequence of the protein in question. These considerations argue that a protein’s “structure” is best described as a distribution over a conformational ensemble consisting of its thermally accessible structures. In this sense, a protein’s conformational ensemble is intimately linked to its function.

Typically, proteins are characterized as being either folded or unfolded. This classification scheme is best understood from an analysis of the corresponding conformational ensembles. Folded proteins have thermally accessible states that are similar to the ensemble average, whereas unfolded (or disordered) proteins sample a relatively vast array of dissimilar conformations during their biological lifetime (2). The native state of a folded protein corresponds to a global energy minimum that is well separated from a panoply of high energy states. The flexibility of a folded protein is related to the width of this minimum energy well (Fig. 1A). Disordered proteins, by contrast, have energy surfaces that contain many local energy minima that are separated by low barriers (on the order of $k_B T$), thereby ensuring rapid transition between structurally dissimilar states during the protein’s biological lifetime (Fig. 1B). The result is a heterogeneous ensemble of thermally accessible conformations.

Although the terms folded and unfolded provide a useful framework for everyday discourse among specialists, classifying proteins in this way does not capture the rich and beautiful complexity that underlies protein structure (3–5). Indeed, a more accurate description would entail a characterization of a protein’s conformational ensemble. The importance of this realization is highlighted by the fact that not all folded proteins are created equal; i.e., some “folded” ensembles are more heterogeneous than others. Similarly, disordered proteins may have ensembles that exhibit preferences for particular structural features. These considerations reinforce the notion that quantitative metrics describing the heterogeneity within a protein’s ensemble would provide a more comprehensive assessment of protein structure.

In a prior work, we introduced a continuous order parameter that describes the conformational heterogeneity in a structural ensemble using a quantitative metric of structural dissimilarity (3). The order parameter is 0 when the root mean square deviation (in terms of the Cartesian coordinates) between any two structures in an ensemble is infinite; i.e., each structure in the ensemble is infinitely different from every other structure. Conversely, the parameter is 1 when each structure in the ensemble is identical to every other structure. Although these upper and lower bounds are clearly theoretical, the notion is instructive. Proteins that are disordered have values close to 0, and those that are folded have values close to 1. Although a protein that is classified as being disordered necessarily falls on the low end of the disorder-order spectrum, it may very well be more ordered than another similarly classified protein.

An Order-Structure Continuum for Describing Protein Structure

Proteins that fall on the low end of the spectrum, by definition, sample a vast set of dissimilar structures in solution. However, such systems are not necessarily devoid of any secondary structural preferences, and therefore are not “unstructured.” This point is important because the nomenclature in the literature with respect to disordered proteins can be confusing. In our view, the term “unstructured” should be used to indicate a lack of regular secondary structure (e.g., $\alpha$-helix or $\beta$-strand), whereas the term “disordered” should indicate a high level of conformational heterogeneity in the conformational ensemble. The fact that proteins can be ordered and yet be devoid of helical or $\beta$-strand structure speaks to the fact that these notions are not the same (6–8).

Disordered proteins can contain regions of local and/or transient residual secondary structure despite a lack of tertiary structure. Indeed, a number of disordered proteins fold into stable conformations upon binding their partners, and residually structured regions may act as binding sites that nucleate...
The abbreviations used are: IDP, intrinsically disordered protein; SAXS, small angle x-ray scattering; TBE, TolB-binding epitope; CBP, CREB-binding protein; CREB, cAMP-response element-binding protein; NCBD, nuclear co-activator-binding domain; IUTD, intrinsically unstructured region within the T-domain; T-domain, translocation domain; R-domain, receptor-binding domain; C-domain, C-terminal cytotoxic domain.

MINIREVIEW: Proteins on the Order-Structure Continuum

Although the axes in this continuum are qualitative, quantitative metrics have been proposed to quantify ensemble heterogeneity, either using atomistic representations for the structural ensemble (3, 18), or from a topological analysis of the energy surface (19). Additionally, secondary structure content can be computed as the percentage of amino acids in a protein with determined or predicted secondary structure, using a variety of metrics. Such metrics form a quantitative basis for a continuous classification scheme for protein structure.

Before moving on, we note that the location of a protein on this three-dimensional continuum has consequences for how that protein can be studied. Experimentally determined crystal structures of proteins correspond to an ensemble average. Because the conformational ensemble of folded proteins (low heterogeneity) contains structures that are similar to their average conformation, crystallographic structures of folded proteins provide great insight into the structural ensemble itself, and that protein’s function. By contrast, crystallizing a disordered protein is not possible precisely because the process of crystallization requires the protein in question to adopt similar structures within the crystal environment (20–22). Structural modeling has therefore played an essential role in the study of very disordered systems (20, 23, 24). Dynamical simulations, for example, combined with restraints derived from NMR and/or SAXS experiments can be used to model important structural features of these proteins (21, 25, 26).

Moving forward, we illustrate how proteins at the far end of the order-structure continuum accomplish a variety of different functions. The examples presented below are by no means intended to be all-inclusive. Nevertheless, they do describe novel and interesting ways that proteins use disorder to accomplish tasks that would be difficult to perform without the considerable flexibility that disorder imparts.

Colicin E9

Colicins are a class of proteins produced by some *Escherichia coli* strains that provide a mechanism for bacteria to compete against similar or related strains when limited environmental resources are available (27, 28). After binding receptors on the surface of the outer membrane of the target cell, colicins are transported via specific translocators on the cell surface to the periplasmic space (27). Colicins then promote bacterial death either via destruction of important components of the peptidoglycan wall, via pore formation in the inner membrane, or by cleaving nucleic acids in the cytoplasmic space (27). Conformational disorder plays a crucial role in ensuring efficient translocation through target cell membranes.

Colicins all have a common structure consisting of an N-terminal translocation domain (T), a receptor-binding domain (R), and a C-terminal cytotoxic domain (C) (27). After binding their cognate outer-membrane receptors, colicins recruit and assemble a translocon, a protein complex that facilitates translocation through the outer membrane surface (Fig. 2). Two proteins that are central components of the translocon are OmpF, Colicins are a class of proteins produced by some *Escherichia coli* strains that provide a mechanism for bacteria to compete against similar or related strains when limited environmental resources are available (27, 28). After binding receptors on the surface of the outer membrane of the target cell, colicins are transported via specific translocators on the cell surface to the periplasmic space (27). Colicins then promote bacterial death either via destruction of important components of the peptidoglycan wall, via pore formation in the inner membrane, or by cleaving nucleic acids in the cytoplasmic space (27). Conformational disorder plays a crucial role in ensuring efficient translocation through target cell membranes.

Colicins all have a common structure consisting of an N-terminal translocation domain (T), a receptor-binding domain (R), and a C-terminal cytotoxic domain (C) (27). After binding their cognate outer-membrane receptors, colicins recruit and assemble a translocon, a protein complex that facilitates translocation through the outer membrane surface (Fig. 2). Two proteins that are central components of the translocon are OmpF,
which is found in the outer membrane, and TolB, which is found in the periplasmic space. Therefore, proper translocon assembly requires the extracellular protein, colicin, to recruit proteins in the outer membrane as well as periplasmic proteins, which are found on the other side of the outer membrane.

The role of disorder in colicin translocation has been best studied for the colicin ColE9 (29). Although both the R-domains and the C-domains of ColE9 are folded, NMR studies of the ColE9 T-domain suggest that its 83-residue N-terminal region is disordered and contains little to no appreciable propensity for secondary structure, findings that place this region at the very far end of the order-structure continuum (29). The R-domain of ColE9 forms a long coiled-coiled structure that binds the outer-membrane receptor, BtuB, of the target cell. When bound to BtuB, ColE9 forms an acute angle with the bacterial outer membrane in a manner such that the disordered region of the T-domain is projected off the outer membrane (Fig. 2A). In this orientation, the disordered region of the T-domain is optimally positioned to recruit, or seek out and bind, the translocator protein, OmpF (30).

The colicin receptor, BtuB, slowly diffuses laterally along the outer membrane, providing ColE9 with a vehicle for its search (31). A useful analogy is to view BtuB as a fishing boat, OmpF as fish, the R-domain of ColE9 as a fishing rod, and the disordered region of the T-domain as the fishing line (Fig. 2A) (32). The disordered T-domain region provides a search radius of ~300 Å, which is centered at the end of the R-domain (33), and its rapid sampling of conformations allows ColE9 to search a much broader surface area for OmpF than would be possible through diffusion alone. Thus, the disordered region of the T-domain provides the same benefits for colicin’s search for OmpF as recasting lines provides to anglers looking for fish. Indeed, early theoretical studies of the role of disorder in molecular recognition suggest that unfolding increases the effective capture
radius of the molecule, thereby facilitating the fast formation of relatively weak contacts at large distances (34). Subsequent studies suggest that unfolded proteins need fewer encounter events (relative to folded proteins) to form stable complexes, and that this explains the relatively fast kinetic rates associated with disordered proteins (35). These observations argue that inherent disorder in the T-domain helps to ensure that OmpF is recruited quickly to ensure efficient translocation of ColE9 (36).

After binding OmpF, the T-domain must pass through the narrow porin to reach the periplasm. OmpF is a trimer containing three identical pores small enough to allow at most a 600-Da molecule to pass through at a time. Because of its intrinsic disorder, the 9-kDa disordered region of the T-domain is able to weave through a pore of OmpF into the periplasm (Fig. 2B). Moreover, the disordered region of the T-domain contains two OmpF interaction sites, residues 2–18 (OBS1) and 54–63 (OBS2), and a TolB-binding site, residues 32–47 (TBE), that facilitate pore entry (33, 37). As the OmpF pores are negatively charged, and OBS1 and OBS2 are positively charged, it has been proposed that an electrostatic interaction drives the entry of OBS1 into the pore and the formation of the initial complex between OBS1 and OmpF (33). Although the subsequent steps in pore navigation are not well understood, eventually the TBE site finds its way into the periplasmic space and to TolB (37). In later stages, OBS2 forms a lower affinity interaction with the initial pore using the same binding site as OBS1 initially used (33), and OBS1 winds back into a different pore on OmpF, likely limiting the movement of the now periplasmic TBE (Fig. 2C) (38).

Whether the TBE binds TolB before or after OBS1 winds back into OmpF is not yet known. However, it has been shown that ColE9 is more lethal to target cells when both OBS1 and OBS2 are present, indicating that the presence of both interactions with OmpF likely stabilizes the contact between the TBE and TolB, thereby also helping to stabilize the complex between TolA and TolB (38). Without disorder, it would not be possible for the T-domain to thread through the porin, bind the periplasmic protein TolB, and then re-enter the OmpF to stabilize the complex. The upshot is that the disordered T-domain helps to ensure that proteins on the periplasmic side of the outer membrane are stabilized in a position that is optimal for colicin translocation.

The interaction of colicin with TolA triggers the protein motive force, which is thought to drive subsequent unfolding of ColE9. The remaining steps of colicin C entry into the cytoplasm are not understood. It is clear, however, that disorder plays a role in the initial steps of ColE9 entry, and similar pathways are thought to be used by related colicins for cell entry.

**4E-binding Protein 2**

The 4E-binding protein 2 (4E-BP2; also known as PHAS-II, phosphorylated heat and acid stable protein regulated by insulin 2 (39)) protein acts as a switch to regulate translation and is critical for development and growth across all cell types (40). This protein converts between an unphosphorylated state, which inhibits translation, and a phosphorylated state, which allows translation (41). NMR studies have shown that this 120-residue protein is disordered in its unphosphorylated state (40).

The unphosphorylated protein binds to and inhibits the eukaryotic initiation factor 4E, eIF4E, which is responsible for initiating translation during cell growth (42). 4E-BP2 competes with the scaffolding protein eIF4G for binding eIF4E; eIF4G promotes translation by facilitating the assembly of translation machinery, whereas 4E-BP2 inhibits translation by sterically blocking the binding site on eIF4E for eIF4G (43). Specifically, eIF4G and 4E-BP2 have a common 7-residue primary sequence motif (YXXXXXlΦ, where Φ represents a hydrophobic amino acid and Χ represents any amino acid) that binds eIF4E (43–45). These competing proteins bind to overlapping sites on eIF4E, so that both cannot bind simultaneously (46, 47). Phosphorylation of 4E-BP2 weakens its binding affinity for eIF4E, which leads to increased binding of eIF4G to eIF4E, and an increased rate of translation (Fig. 3).

Analyses of NMR chemical shifts indicate that the 7-residue eIF4E-binding site (residues 54–60) in unphosphorylated 4E-BP2 has transient helical structure when not bound to eIF4E, and is flanked by two small segments that have residual extended structure (46). Four other small regions (residues 1–5, 33–37, 86–89, and 96–105) are also predicted to have transient helicity (46). NMR heteronuclear NOEs are consistent with these observations, as their values are higher in these regions than expected for entirely disordered proteins, indicating transient local or tertiary structure (46). Thus, unphosphorylated 4E-BP2 has low order, but higher levels of secondary structure than the discussed region of ColE9, placing it farther from the origin in the helical content and β-strand content directions of the order-structure continuum.

Upon binding eIF4E, residues 54–60 in 4E-BP2 fold into an α-helix and residues 78–82 also form a transient interaction with eIF4E (46). NMR studies show that the rest of 4E-BP2 remains disordered in the eIF4E-bound state, and some regions actually become more disordered upon binding (40, 46). Despite remaining disordered, some changes in chemical shifts upon binding eIF4E are observed (46), and SAXS experiments suggest that 4E-BP2 becomes more compact upon binding (48).

Extracellular signaling for phosphorylation of 4E-BP2 by cellular kinases, e.g., through growth factors or mitogens, leads to a reduction in affinity of 4E-BP2 for eIF4E (41, 42, 49). Thus, phosphorylation disrupts eIF4E binding to 4E-BP2, thereby allowing eIF4G to bind free eIF4E, a process that promotes translation. 4E-BP2 is phosphorylated initially on residues Thr37 and Thr46 and later on residues Ser65 and Thr70 (49). The initial phosphorylation on Thr37 and Thr46 greatly weakens the interaction between 4E-BP2 and eIF4E, and subsequent phosphorylation of residues Ser65, Thr70, and Ser83 further lowers the affinity between 4E-BP2 and eIF4E (41, 49). Accompanying studies suggest that 4E-BP2 forms a four-strand β-structure upon phosphorylation of residues Thr37 and Thr46, and that this structure is further stabilized by the phosphorylation of Ser65, Thr70, and Thr83 (41). Residues 54–56, which form part of a helix when bound to eIF4E, are incorporated in one of the β-strands, and residues 58–60 form a disordered loop (Fig. 3) (41). Thus, phosphorylation of the disordered 4E-BP2 triggers its transformation into a conformation that is unfavorable for binding eIF4E (Fig. 3).
Although many IDPs fold into a stable complex upon binding their partners, 4E-BP2 is an example of a protein that remains largely disordered in its bound state, and this disorder is crucial to its function. The disorder of 4E-BP2 in both its unbound and eIF4E-bound state allows its phosphorylation sites to remain exposed (41, 46, 49). Phosphorylation of 4E-BP2 causes 4E-BP2 to fold into a structure, in which one of the β-strands involves residues that form a helix when bound to eIF4E (41). The release of 4E-BP2 from eIF4E upon phosphorylation of 4E-BP2 is likely due to this conformational change, and thus phosphorylation can result in the rapid release of 4E-BP2 from eIF4E, permitting initiation of translation. In this way, the intrinsic disorder of 4E-BP2 when bound permits it to react quickly to hormonal signals calling for the initiation of translation.

**NCBD**

The nuclear co-activator-binding domain (NCBD) of the CREB-binding protein (CBP) is a transcriptional co-activator. This 59-residue domain within CBP (residues 2058–2116 of human CBP) interacts with a diverse set of proteins, including transcription factors and various elements of the transcriptional machinery (50, 51). Several experimental observations suggest that NCBD has poorly dispersed chemical shifts and weak long-range NOEs, features associated with a lack of stable tertiary structure (52–54). Despite this, circular dichroism spectra suggest that NCBD retains significant helical content (50). Given NCBD’s high degree of native secondary structure coupled with its lack of a stable fold, this protein has been classified as a molten globule (52, 53, 55). A number of studies, however, suggest that the situation is likely more complex. NCBD has a hydrophobic core that has a sigmoidal unfolding curve in the presence of urea, and NMR relaxation data argue that it slowly interconverts between several conformations on the NMR time scale (54, 56). Unlike traditional IDPs that rapidly fluctuate between dissimilar conformations corresponding to local energy minima separated by low barriers, NCBD samples states separated by relatively large barriers, leading to longer transition times. In our parlance, the fact that NCBD samples distinct conformational states in solution on the millisecond time scale places it in the low order (relative to archetypal folded proteins) and high structure region of the order-structure continuum.

Two models for the hydrophobic core of NCBD have been proposed, NCBD-1 (54) and NCBD-2 (50), where the structure of NCBD-1 corresponds to a more highly populated conformer in the unfolded ensemble. In both models, the NCBD core contains three helices, whose orientations and lengths vary depending on the identity of their binding partners (Fig. 4) (55). NMR studies of unbound NCBD suggest that the protein fluctuates on a millisecond time scale between two conformational states, including a dominant state similar to conformation NCBD-1, in which helices 1 and 2 form contacts, and a less prevalent state in which this contact is replaced by interactions between helix 1 and 3, more similar to the conformation bound to interferon regulatory factor 3 (IRF-3) (56, 57) (Fig. 4). Molecular dynamics simulations of the unbound state have shown that NCBD samples a wide variety of conformations characterized by different orientations and lengths of its helices (58, 59). Long time-scale simulations indicated that NCBD samples conformations similar to each known bound structure at low rates. These simulations argue that the IRF-3 bound conformation is only rarely accessible from the unbound state, indicating that the presence of that binding partner may be necessary for
MINIREVIEW: Proteins on the Order-Structure Continuum

that illustrate how disorder can expand the repertoire of protein functions.

In addition to playing an important role in many biochemical processes, disordered proteins have been implicated in a number of diseases, both as pathogens and as chaperones (60–66). Thus, a better understanding of these proteins may provide a platform for the engineering of novel therapeutic agents (67, 68). More generally, an improved understanding of the relationship between a protein’s primary sequence and its structural ensemble is essential for the design of novel proteins that could be used in technology and medicine. A crucial step in this direction involves an expansion in methods for studying the energetic landscape of proteins. Recent adaptions to crystallography and NMR are providing insight into partially stable molecular states, thus providing new glimpses into protein ensembles, albeit for proteins that lie in the more ordered region of the continuum (69, 70).

Our discussion has been framed within the context of an order-structure continuum for protein structure classification. In doing so, we strive to reinforce the realization that a binary classification of proteins as either “folded” or “unfolded” does not capture the wide variety of flexible architectures available to proteins (4, 5, 13). Our order-structure continuum provides a qualitative overview of the varying degree of order and secondary structure content among proteins, and here we have discussed examples of functions carried out by proteins that have varying degrees of order and secondary structure. Going forward, experimental methods to quantify the amount of disorder in a protein, especially under various physiological conditions, would lead to a better classification system, provide important grounds for insight into how a protein’s flexibility enables its function, and guide the design of further experiments to study that protein’s conformational ensemble. Overall, the ability to relate a protein’s sequence to its conformational ensemble and its range of functions would enable exciting advances in biomedicine and bioengineering.

References


NCBD to sample that state (58). A separate molecular dynamics simulation study of NCBD binding to its interaction domain on the human activator for thyroid hormone and retinoid receptors protein (ACTR) or IRF-3 indicated that NCBD samples its bound conformation for each of these partners much more readily in the presence of that partner than in its absence (59).

NCBD is a hub domain of CBP that interacts with many binding partners, and is known to bind different partners using different arrangements of its helices (Fig. 4). Overall, the ability of NCBD to sample different helical lengths and orientations enables it to bind a repertoire of distinct binding partners. The intrinsic low order facilitates these interactions, and its high secondary structure content ensures that it does not incur the significant entropic losses that are associated with having to adopt a folded state upon binding.

Conclusion

Our understanding of protein disorder and the role that it plays in protein function has blossomed over the past several decades. Knowledge of the ways in which disorder can add to the rich complexity of proteins has evolved for a number of proteins, and the growth of research in this area ensures that the rate of progress in this burgeoning field will only increase. In this minireview, we have attempted to highlight a few examples...
between the p160 coactivator and CREB-binding protein. *Protein Sci.* **13**, 203–210