Conformational Ensembles Exhibit Extensive Molecular Recognition Features
Robert I. Cukier

Department of Chemistry, Michigan State University, 578 S. Shaw Lane, East Lansing, Michigan 48824-1322, United States

ABSTRACT: Intrinsically disordered proteins (IDPs) are important for signaling and regulatory pathways. In contrast to folded proteins, they sample a diverse conformational space. IDPs have residue ranges within a sequence that have been referred to as molecular recognition features (MoRFs). A MoRF can be viewed as contiguous residues exhibiting a conformational disorder that become ordered upon binding to another protein or ligand. In this work, we introduce a structural characterization of MoRFs based on entropy and mutual information (MI). In this view, a MoRF is a set of contiguous residues that exhibit a large entropy (from rotameric residue sampling) and large MI, the latter indicating a dependence among the residues’ rotameric sampling comprising the MoRF. The methodology is first applied to a number of ubiquitin ensembles that were obtained based on nuclear magnetic resonance experiments. One is a denatured Ub ensemble that has a large entropy for various unitSizes (number of contiguous residues) but essentially zero MI, indicating no dependence among the residue rotamer sampling. Another ensemble does exhibit extensive regions along the sequence where there are MoRFs centered on nonsecondary structure regions. The MoRFs are present for unitSizes 2−10. That a substantial number of MoRFs are present in Ub strongly suggests a conformational selection mechanism for this protein. Two additional ensembles for the cyclin-dependent kinase inhibitor Sic1 and for the amyloid protein α-synuclein, which have been shown to be IDPs, are also analyzed. Both exhibit MoRF-like character.

1. INTRODUCTION

Although folded proteins carry out enzymatic reactions, intrinsically disordered proteins (IDPs) are key for the regulation of transcription and translation and, more generally, for cell signaling.1−10 Protein−protein and protein−ligand interactions often take place via a disorder-to-order transition, whereby an unstructured free protein becomes structured by interaction with a partner protein or ligand, forming a functional interface by a decrease in the interfacial free energy. In a conformational selection mechanism, the presence of special conformations of an IDP may be required to bind a partner protein or ligand. In an induced fit mechanism, the IDP becomes structured upon interaction with its binding partner.4,11−13 The timing and ordering, of what is in essence a coupled binding and folding process, is crucial.3,14−18

An aspect of IDPs that has garnered attention is the identification of limited-in-extent residue subsequences that exhibit some disorder, before binding orders them. These regions of a sequence have been termed as molecular recognition features (MoRFs) and short linear peptide motifs,19 the latter more associated with shorter subsequences that make contact with a binding partner.19−21 For specificity, we will use the MoRF terminology. An IDP can be thought of as a MoRF subsequence interspersed within more rigid subsequences, though the disorder can certainly extend over the entire sequence.

The most predictive work on MoRFs is based on a sequence without account of structural information.6,21−24 For a given sequence of a sequence, these methods use residue properties such as hydrophobicity and charge and their distribution in the subsequences to predict regions that are likely to be disordered. Here, rather than predicting which subsequences are disordered, we analyze conformational ensembles for a given sequence and, from that information, assess whether various subsequences can be viewed as MoRFs. This approach of course has the drawback that structural information is much more limited than the number of subsequences that are known, but the structural information does allow for a more precise characterization of the amount of order.

Obtaining conformational ensembles for IDPs from both experiment and computation is a difficult task. Various methodologies such as nuclear magnetic resonance (NMR) and small-angle X-ray scattering (SAXS) are currently the most
employed methods, but these methods may suffer from under determination and often rely on computer-generated ensembles to refine the NMR data or other data. From the computational perspective, in principle, molecular dynamics (MD) can provide a sufficiently long trajectory that samples from a Boltzmann equilibrium conformational ensemble. However, as much as it is well-appreciated, MD on complex systems, such as proteins, is limited by having to overcome barriers in a rugged energy landscape. Furthermore, as with any MD-based study, there is the caution of coming to conclusions based on the particular force field used. Both issues are of particular concern for IDP simulations.

In this work, we develop structure-based methods to quantify the degree of order in subsequences of a given sequence. Do the residues of the subsequence "randomly" sample, for each residue, their possible conformations independently of the other residues, or do they exhibit an internal order, whereby the conformation of some residues determines that of other residues? If the latter, we consider such a subsequence to be a MoRF. Conformational states will be defined by backbone $\phi$ and $\psi$ dihedral angles obtained for each residue from Ramachandran space sampling in an ensemble of conformers. Mutual information (MI), based on information entropy, will be the metric that provides a quantitative measure of the dependence between residue dihedral sampling of rotamers. In this view, a MoRF is a unit of contiguous residues that has a relatively large Ramachandran entropy and large MI, indicating that the residues of the MoRF are sampling distinct rotamers but with a strong dependence among these residues’ conformational sampling. The implication is that there are distinct conformers being sampled, yet the populations of these conformers are quite limited. Mechanistically, the presence of MoRFs favors a conformational selection mechanism. On the other hand, residues with a large entropy but small MI are essentially random sampling conformations and will be classified as not-MoRF subsequences.

This structural approach to MoRFs is first applied to ubiquitin (Ub). One role of Ub is protein degradation, accomplished by its ability to recognize a variety of ubiquitin (Ub). One role of Ub is protein degradation, accomplished by its ability to recognize a variety of ubiquitin (Ub). Another role of Ub is control of protein synthesis. By controlling the degradation of certain proteins, Ub helps to regulate the activity of enzymes and other proteins in the cell. A MoRF analysis can provide insights into the structural and functional importance of different regions of Ub.

Two other proteins, the cyclin-dependent kinase (CDK) inhibitor Sic1 and the amyloid protein $\alpha$-synuclein ($\alpha$Syn), are also analyzed for MoRF character. Both were investigated by NMR and SAXS methodologies, and ensembles of conformers were generated. On the basis of the generated structures, they sample a very diverse set of conformers. Our analysis supports this conclusion.

The plan of the remainder of this paper is as follows. Section 2 assesses the four Ub ensembles that were nominally suitable to this study and shows that two are appropriate for our purposes, having a sufficient diversity of distinct, rotameric Ramachandran space conformations. Section 3 presents the Sic1 and $\alpha$Syn analyses. Section 4 outlines the main conclusions of this study and some pitfalls and prospects of a structural MoRF analysis. Section 5 describes our methodology.

2. UB RESULTS

2.1. The UB_ERIDU Ensemble. The NMR data for this ensemble were obtained under denaturing conditions of 8 M urea and pH 2. From the 1000 structures given as pdb files, Rama(1) plots and MoRF data were evaluated as outlined in the Computational Methods Section. The ensemble refinement of intrinsically disordered and unstructured molecules (ERIDU) Rama(1) data are displayed in Figure 1. The pale blue color is Gly. Otherwise, the residues are mainly sampling $\alpha$ ($-90,-30$), $\beta$-PPII ($-120,150$), and $\alpha$L ($90,45$) regions.

![Figure 1. Residue Ramachandran plots for the UB_ERIDU-denatured ensemble (by residue type). Gly is light blue. The other residue main populations are in the $\alpha$ ($-90,-30$), $\beta$-PPII ($-120,150$), and $\alpha$L ($90,45$) regions.](image-url)
nuclear Overhauser effects (NOEs) data along with the restraints (EROS) ensemble and from the ligand-bound Ub X-ray structures. The regions of high root mean square fluctuation (RMSF), as obtained in their ensemble refinement with orientational restraints (EROS) ensemble and from the ligand-bound Ub X-ray structures. The regions of large RMSF found that there are approximatively 1 residues 5–15, region 2 residues 30–40, 3 residues 45–52, and tail residues. Comparison with Table 1 identifies MoRFs in regions 1, 2, and 3, along with some not-MoRF contributions because of our low-entropy cutoff. Thus, the agreement with the experimental data regarding regions of high entropy is robust. What, of course, cannot be assessed from RMSF measures alone, that only indicate independent residue entropies, is the correlation among residues that lead to the concept of a MoRF.

There are potential functional interpretations of our categorization. The MoRF that covers approximately residues T7–T12 corresponds to the β1–β2 loop region (Figure 3) that is known to be a molecular recognition region. The combination of large entropy (Figure 5) and MI (Table 1), thus indicating flexible but correlated residues, may be significant for molecular recognition. The MoRF spanning Q40-L50 spans the hydrophobic patch centered on I44. This

MI plots for unitSizes 2, 4, 8, and 10 are displayed in Figure 4. There are many units whose MInormed values are essentially unity (complete dependence), but not all have corresponding substantial entropy, as made quantitative below. Thus, there are a limited number of MoRFs (large entropy and large MInormed). Note that even as the unitSize increases to 10, there are still quite a few MoRFs—the dependence between the two halves of the unit is substantial and therefore the dependence spans extended sets of contiguous residues. Units with a large entropy but small MInormed values in contrast are those that are sampling rotameric states but in an essentially independent fashion. These are classified as not-MoRF units. To make the term substantial entropy quantitative, the entropy of each unit on a per residue basis, assuming each residues’ rotameric sampling is independent, is plotted in Figure 5. Candidate MoRFs have substantial entropy per residue on the scale of ln(2) = 0.69 for four rotamers with two 0.5 populations and ln(4) = 1.39 for four rotamers with an equal population.

Table 1 provides an overview of some of the MoRFs and not-MoRFs that are present for the different unitSizes, along with not-IDP ranges that do not sample multirotameric conformations. Included are the residue ranges and the secondary DSSP designations, as displayed in Figure 3.

The above categorization is in good agreement with the DSSP structural classification of UB, as shown in Figure 3, where one associates the secondary structure with more constrained and loop regions with the more fluctuating regions of a sequence. It is also in good agreement with the NMR data. The Lange et al. NMR experiment (their Figure 1C) provides regions of high root mean square fluctuation (RMSF), as obtained in their ensemble refinement with orientational restraints (EROS) ensemble and from the ligand-bound Ub X-ray structures. The regions of large RMSF found that there are approximately region 1 residues 5–15, region 2 residues 30–40, region 3 residues 45–52, and tail residues. Comparison with Table 1 identifies MoRFs in regions 1, 2, and 3, along with some not-MoRF contributions because of our low-entropy cutoff. Thus, the agreement with the experimental data regarding regions of high entropy is robust. What, of course, cannot be assessed from RMSF measures alone, that only indicate independent residue entropies, is the correlation among residues that lead to the concept of a MoRF.

There are potential functional interpretations of our categorization. The MoRF that covers approximately residues T7–T12 corresponds to the β1–β2 loop region (Figure 3) that is known to be a molecular recognition region. The combination of large entropy (Figure 5) and MI (Table 1), thus indicating flexible but correlated residues, may be significant for molecular recognition. The MoRF spanning Q40-L50 spans the hydrophobic patch centered on I44. This

0.5) = 0.69; Hind(0.25/0.75) = 0.56; Hind(0.15/0.85) = 0.42; and Hind(0.1/0.9) = 0.32. From the values of Hind, the sum of the Rama(1) entropies over each two- and each four-residue neighbor stretch, there is rotameric sampling in some or all of the residues involved, but the corresponding MIs are negligible. When scaled to the bounded values 0 ≤ MInormed(X;Y) < 1, the values are all around zero. Thus, the rotamer samplings of the two half units of a unit are independent. This result is a good check on the methodology, as one would expect a denatured ensemble to produce, for the various unitSizes, a large entropy but negligible dependence among their parts.

2.2. UB_2K39 Ensemble. The UB_2K39 ensemble was obtained using microsecond NMR RDCs supplemented by nuclear Overhauser effects (NOEs) data along with the CONCOORD program to generate 116 structures. As noted by Lange et al., an intriguing result is that thisapo Ub ensemble spans the Ub conformations found by crystallography for 46 Ub complexes. The longer time span of the ensemble spans the Ub conformations found by crystallography for 46 Ub complexes. The longer time span of the experiment was suggested as the origin of the diverse ensemble.

Figure 2. UB_ERIDU ensemble entropy and normed MI, MInormed, plots for unitSizes two [Rama(2) space] and four [Rama(4) space]. Hind is the entropy summed over, respectively, two and four contiguous residues in the sequence. MInormed is the corresponding MI bound defined in eqs 5.5 and 5.6. It is clear from these plots that this ensemble is essentially sampling “random” rotameric Ramachandran states, as appropriate to a denatured ensemble.

Figure 3. DSSP secondary structure assignment for Ub (PDB 1UBQ).

Table 1 provides an overview of some of the MoRFs and not-MoRFs that are present for the different unitSizes, along with not-IDP ranges that do not sample multirotameric conformations. Included are the residue ranges and the secondary DSSP designations, as displayed in Figure 3.

The above categorization is in good agreement with the DSSP structural classification of UB, as shown in Figure 3, where one associates the secondary structure with more constrained and loop regions with the more fluctuating regions of a sequence. It is also in good agreement with the NMR data. The Lange et al. NMR experiment (their Figure 1C) provides regions of high root mean square fluctuation (RMSF), as obtained in their ensemble refinement with orientational restraints (EROS) ensemble and from the ligand-bound Ub X-ray structures. The regions of large RMSF found that there are approximately region 1 residues 5–15, region 2 residues 30–40, region 3 residues 45–52, and tail residues. Comparison with Table 1 identifies MoRFs in regions 1, 2, and 3, along with some not-MoRF contributions because of our low-entropy cutoff. Thus, the agreement with the experimental data regarding regions of high entropy is robust. What, of course, cannot be assessed from RMSF measures alone, that only indicate independent residue entropies, is the correlation among residues that lead to the concept of a MoRF.

There are potential functional interpretations of our categorization. The MoRF that covers approximately residues T7–T12 corresponds to the β1–β2 loop region (Figure 3) that is known to be a molecular recognition region. The combination of large entropy (Figure 5) and MI (Table 1), thus indicating flexible but correlated residues, may be significant for molecular recognition. The MoRF spanning Q40-L50 spans the hydrophobic patch centered on I44. This

Figure 2. UB_ERIDU ensemble entropy and normed MI, MInormed, plots for unitSizes two [Rama(2) space] and four [Rama(4) space]. Hind is the entropy summed over, respectively, two and four contiguous residues in the sequence. MInormed is the corresponding MI bound defined in eqs 5.5 and 5.6. It is clear from these plots that this ensemble is essentially sampling “random” rotameric Ramachandran states, as appropriate to a denatured ensemble.

0.5) = 0.69; Hind(0.25/0.75) = 0.56; Hind(0.15/0.85) = 0.42; and Hind(0.1/0.9) = 0.32. From the values of Hind, the sum of the Rama(1) entropies over each two- and each four-residue neighbor stretch, there is rotameric sampling in some or all of the residues involved, but the corresponding MIs are negligible. When scaled to the bounded values 0 ≤ MInormed(X;Y) < 1, the values are all around zero. Thus, the rotamer samplings of the two half units of a unit are independent. This result is a good check on the methodology, as one would expect a denatured ensemble to produce, for the various unitSizes, a large entropy but negligible dependence among their parts.

2.2. UB_2K39 Ensemble. The UB_2K39 ensemble was obtained using microsecond NMR RDCs supplemented by nuclear Overhauser effects (NOEs) data along with the CONCOORD program to generate 116 structures. As noted by Lange et al., an intriguing result is that this apo Ub ensemble spans the Ub conformations found by crystallography for 46 Ub complexes. The longer time span of the experiment was suggested as the origin of the diverse ensemble for free Ub.

The Dictionary of Protein Secondary Structure (DSSP) secondary structure assignment for Ub (PDB 1UBQ) is given in Figure 3.

Figure 3. DSSP secondary structure assignment for Ub (PDB 1UBQ).
Ub-binding domain shows maximal contact with binding proteins. Lange et al. note that there is large RMSF in this domain that is correlated with many binding protein contacts.

The combination of large entropy and MI found in this region may enhance the ability to bind partner proteins. The not-MoRF found around residues 65–74 that is mainly β-strand.

Figure 4. Values of the total entropy assuming independence of the two half units, X and Y, and the normalized MI (0 ≤ MI_{normed}(X,Y) ≤ 1) for units of the indicated sizes. Those units with a large entropy and large MI are the MoRFs and those with a large entropy but small MI are the not-MoRFs. The plots indicate that there are MoRFs even with increasing unitSize; dependence can extend over a number of contiguous residues. The numbering of the plots is from the middle residue of the unit, that is, for unitSize 10, the residue pointed to in the residue numbering is the middle residue of the unit. For example, the first unit, QIFVKTLTGK, starts at residue 6, which is the K in the middle of this unit.

Figure 5. Total independent entropy $H_{\text{ind}}^k = \sum_{p=1}^{P-1} H(X_{k+p})/P$ per residue of unit $k$ for unitSizes $P = 2, 4, 8, \text{ and } 10$. High entropy is concentrated in nonsecondary structure regions of the sequence, as defined in Figure 3.

Ub-binding domain shows maximal contact with binding proteins. Lange et al. note that there is large RMSF in this domain that is correlated with many binding protein contacts.

The combination of large entropy and MI found in this region may enhance the ability to bind partner proteins. The not-MoRF found around residues 65–74 that is mainly β-strand.
Table 1. Some of the MoRFs and Not-MoRFs for Various UnitSizes with Their DSSP Secondary Structure

| class       | unitSize | unit                  | MoRF<sup>a</sup>          | not-MoRF<sup>b</sup>          | not-IDP<sup>c</sup>          |
|-------------|----------|-----------------------|---------------------------|------------------------------|-------------------------------|
| class       |          |                       | TG(9–10)                  | GK(10–11)                    | 20–40                        |
|             | 2        |                       | LTGK(8–11) turn-bend      | GTKT(9–12)                   | 20–40                         |
|             | 4        |                       | AG(46–47)                 | TTLT(12–15) β-strand         | 20–40                         |
|             |          |                       | DG(52–53)                 | IFAG(44–47) β-strand-turn    | 20–40                         |
|             |          |                       | QK(62–63)                 | FAGK(45–48)                  | 20–40                         |
|             |          |                       | LEDG(50–54) unassigned-turn | QKES(62–65) unassigned-turn | 20–40                         |
| not-MoRF<sup>b</sup> |          |                       | GK(10–11)                 | GKT(10–13) unassigned-β-strand | 20–40                         |
|             |          |                       | RL(42–43)                 | LIFA(43–47) β-strand         | 20–40                         |
| not-IDP<sup>c</sup> |          |                       | 20–40                     | 20–40                         | 20–40                         |

<sup>a</sup>MoRF: large Hind and large MInormed. <sup>b</sup>Not-MoRF: large Hind and small MInormed. <sup>c</sup>Not-IDP: small Hind and small or large MInormed. "The assignment is on the following line of the unit.

has a large entropy but small MI. This region is peripheral to the hydrophobic patch. The proteins here have a different character, being mainly α helical. Finally, the region spanning residues 20–40 that are α helical in the DSSP classification (Figure 3) has a small entropy that defines not-IDP in our classification scheme.

The designation of a unit as MoRF or not-MoRF of course has to be made by picking a dividing value of entropy along with normed MI. Figure 5 displays independent entropies per residue of the units and makes it clear that entropies in the nonsecondary structure parts are comparable to what would be expected for more than one rotamer sampling of individual residues. Within these large entropy regions of the sequence, values of the normed MI that are approaching unity provide a quantitative measure to label as a MoRF. Tables of all MoRFs and not-MoRFs are given in the Supporting Information (Tables 1–4) for unitSizes 2, 4, 8, and 10, along with their definitions. Other choices for the dividing lines between MoRFs and not-MoRFs lead to some changes in the categorization.

2.3. UB_2K39 Tail (Residues 72–75) Conformations. The Ub backbone is relatively rigid with the exception of the C terminus extending over residues 72–76 with sequence RLRRG. A visualization of the UB_2K39 ensemble of 116 structures is shown in Figure 6.

A partitioning of the UB_2K39 ensemble into clusters in the dihedral space may reveal distinct conformations of these C-terminal residues. Because our clustering scheme is based on the Rama(1) residue information, the analysis can only be carried out on residues 72–75. The last unitSize 8 second half-unit spans these residues. Its clustering then partitions the 116 samples into various conformers. Some of the high population clusters are displayed in Figure 7.

Visual examination of these clusters and the corresponding values of the cluster centroids of the φ and ψ dihedral angles for residues 72–75, given in Figure 7 caption, show that the clustering algorithm does a good job of resolving the Lange et al. ensemble into rather distinct contributions. That there are distinct groups of conformations found for these residues does suggest that there may be a conformational selection mechanism in operation for this region of Ub.

3. SIC1 AND αSYN RESULTS

The entropies of the UB_2K39 ensemble show multirotamer sampling but may not sample them as extensively as other IDP candidates. Two proteins, Sic1 and αSyn, that have been shown to be IDPs where conformational ensembles are available for our analysis do show, as found below, larger entropy scales. Thus, it is of interest to consider their MoRF characters.

3.1. Sic1. The CDK inhibitor Sic1 has been studied by NMR and found to be an IDP. It interacts with Cdc4, a subunit of a Ub ligase, that aids in Sic1 degradation via a subunit of a Ub ligase, that aids in Sic1 degradation via a...
ensemble, while characterized by great flexibility, still maintains significant nonrandom character. From the point of view of the current investigation, the large differences in the dihedral sampling found, coupled with a small ensemble, suggest that the sampling may correspond to a large entropy and substantial MI values. The phosphorylation sites of Sic1 are Thr-2, Thr-5, Thr-33, Thr-45, Ser-69, Ser-76, and Ser-80.

**Figure 7.** Structures for three distinct clusters. The \((\phi, \psi)\) angle centroids (in degrees) of residues 72−75 of the clusters are top (79.37, −37.26), (−124.34, 154.15), (−82.49, 121.84), (−122.16, 151.51); middle (−96.31, 174.92), (−94.62, 153.77), (90.57, 122.87), (116.54, 156.99); bottom (−73.97, −27.71), (−135.98, 149.98), (−112.62, −35.68), (115.83, 157.81).

A plot of the \(H_{\text{ind}}\), the entropy of each unit on a per residue basis for unitSize 4, assuming each residues’ rotameric sampling is independent, is shown in **Figure 8**. The scale of entropy per residue [four rotamers with two 0.5 populations give \(\ln(2) = 0.69\) and four rotamers with equal population gives \(\ln(4) = 1.39\)] shown in the figure indicates substantial rotamer sampling for each residue, when they are treated as independent. The values are smaller or moderate around most of the sites of phosphorylation. That may indicate more restrained sets of rotameric sampling around these sites.

The unitSize 4 Hind values for the half-units and the corresponding MI values are shown in **Figure 9**. The Hind values track the \(H_{\text{ind}}\) values in **Figure 8**. Note that our definition of Hind is the entropy of the half-units, here unitSize 2, whereas \(H_{\text{ind}}\) asserts that all residues in a unit are independent samplings and so maximal possible entropy. The conclusion still remains that the entropy around the phosphorylation sites is quite reduced. The MI values support independent. The values are smaller or moderate around most of the sites of phosphorylation. That may indicate more restrained sets of rotameric sampling around these sites.

**Figure 8.** Total independent entropy \(H_{\text{ind}} = \sum_{p=0}^{P-1} H(X_{i+p})/P\) per residue of unit \(k\) for unitSize 4. These entropies tend to be smaller around the phosphorylation sites.

**Figure 9.** Values of the total entropy assuming independence of the two half units, \(X\) and \(Y\), (top plot) and the normalized MI \((0 \leq \text{MI}_{\text{normed}}(X;Y) \leq 1)\) (bottom plot) for unitSize 4. Units with large entropy and MI are the MoRFs and those with large entropy but small MI are the not-MoRFs. The plots indicate that there are MoRFs concentrated around the phosphorylation sites. The numbering of the plots is from the middle residue of the unit, here residue 3.
this assertion in that they are larger around the sites of phosphorylation. Thus, these results are in agreement with the conclusions of Mittag et al.\textsuperscript{61,62} regarding robust conformational sampling but certainly not random coil-like behavior. Furthermore, the above analysis suggests that the sampling is more restrained around the sites of phosphorylation that interact with the Cdc4 receptor.

### 3.2. α-Synuclein

αSyn is a 140 residue, strongly disordered IDP that can aggregate under physiological conditions. Its abnormal aggregation is a hallmark of Parkinson’s disease among other neurodegenerative diseases.\textsuperscript{63,64} Numerous experiments have shown that it is quite disordered,\textsuperscript{64} but when it interacts with a biological membrane, an N-terminal region (residues 1–90) becomes more ordered, whereas the C-terminal region (residues 90–140) remains disordered.\textsuperscript{64,70,71} A division of the sequence into membrane recognition (residues 1–60), aggregation-prone (residues 60–95), and intrinsically disordered (residues 95–140) subsequences has been suggested.\textsuperscript{64} Without membrane interaction, αSyn is overall disordered, as shown by a host of NMR methodologies as well as SAXS.\textsuperscript{53–69} Although αSyn is strongly disordered, experiments suggest that it still is not a random coil, with a propensity to have enhanced β-polyproline II (PPII) and β sheet Ramachandran plane sampling relative to random coils.\textsuperscript{64,67} Hydrodynamic and other measurements suggest that it has a smaller radius of gyration than appropriate for a random coil.\textsuperscript{68,69,72} Interactions among residue ranges in the C and N terminals and middle regions of the sequence have been suggested as sources of subtle interactions that produce nonrandom coil behavior.\textsuperscript{65–69}

Before presenting the MoRF results, it is worth examining the underlying properties of the ensemble that ultimately lead to the MoRF/not-MoRF character. As noted in Section 5.3, the ensemble was constructed by use of NMR PRE measurements as constraints on a high-temperature MD simulation.\textsuperscript{66} The PRE measurements introduce a number of spin labels that, while picked to minimize structural perturbations, still alter the native protein sampling, and the MD of course introduces force field sampling issues that are especially significant for IDPs.\textsuperscript{30–42}

One way to characterize the ensemble is via its propensity to sample different regions of the Ramachandran plane. Two definitions of the Ramachandran plane sampling that are useful for IDPs (see e.g., Figure 1) are (1) α (angles $\varphi < 0$ and $\psi < 0$), β sheet and PPII (angles $\varphi < 0$ and $\psi > 0$), and Gly (angles $\varphi > 0$ and $\psi > 0$). (2) Split of β sheet-PPII into PPII (angles $−100 < \varphi < 0$ and $\psi > 0$) and $β$ ( $\varphi < −100$ and $\psi > 0$). These definitions were made on the basis of a “dot” plot of the Ramachandran plane populations of all residues similar to that shown in Figure 1 for Ub. They were chosen based on the following observations: the PPII and $β$ sheet occupations are not very distinct so they can be combined. The $α$ region is less well-defined than that in Figure 1—instead of the characteristic downward sloping of $\psi$ versus $\varphi$, it is more spread out and does not slope downward. The right side ($\varphi > 0$) occupation corresponds mainly to glycine residue sampling.

On the basis of these Ramachandran regions, Figure 10 displays two views of this sampling where in both the cumulating fractions of the various regions of the Ramachandran plane are presented. That is, for each residue, the ensemble-averaged occupation fraction is obtained, and a running addition of their values is evaluated. If there were a distinction between the N- and C-terminal regions based on different Ramachandran plane occupation propensities, then

**Figure 10.** Cumulating Ramachandran plane population fractions as a function of the residue number. Left panel: $α$ (angles $\varphi < 0$ and $\psi < 0$), $β$ sheet-PPII (angles $\varphi < 0$ and $\psi > 0$), and Gly (angles $\varphi > 0$ and $\psi > 0$). Right panel: split of $β$ sheet-PPII into PPII (angles $−100 < \varphi < 0$ and $\psi > 0$) and $β$ ( $\varphi < −100$ and $\psi > 0$).

**Figure 11.** Total independent entropy $H_{k}^{\text{ind}} = \sum_{p=0}^{k} H(X_{k+p})/P$ per residue of unit $k$ for unitSizes 8 and 12.
that would be evident in the plot. The plots show that there is no distinction for this ensemble of conformers that can be discerned.

The total (over all residues) fractional populations are $\alpha$ total = 0.213, PPII Total = 0.372, $\beta$ total = 0.335, and GlyTotal = 0.080, indicating strong bias toward PPII and $\beta$ sheet regions of the Ramachandran plane that can be taken as an excellent signature of an IDP. These results are in good agreement with previous conclusions$^6$ based on NMR and SAXS experiments and computational methods that also construct ensembles. That work also finds enhanced PPII sampling relative to random coil sampling distributed throughout the sequence.

Figure 10 shows that this enhancement is present in the aggregation-prone subsequence (residues 60−95) that may nucleate aggregation and $\beta$ sheet formation, though there is no “excess” enhancement in this region relative to the rest of the sequence.

Figure 11 displays the total independent entropy

$$H_k^{\text{ind}} = \sum_{p=0}^{p-1} H(X_{k+p})/P$$

per residue of unit $k$ for unitSizes $P = 8$ and $P = 12$. The $P = 16$ result is quite similar to the $P = 12$ result. This entropy treats each residue as independent, sums over all residues in the indicated unitSize, and normalizes the sum with the number of residues. Note that for $P = 12$ and the residue 60 point in the plot, residues 60−71 are included in $H_k^{\text{ind}}$. The plots show a dramatic dependence on the unit composition. In particular, there is a distinctive decrease in the initial part of the aggregation-prone region. That indicates, along with part of the C-terminal and N-terminal subsequences, that the ensemble is very limited in sampling in these regions.

The summed half-unit entropies, $H_{\text{half}}$, and normed MI, $M_{\text{normed}}$, are shown in Figure 12 for unitSizes 4 and 8. They show that the Mnormed values are very small for unitSize 4 and larger but still small for unitSize 8, in spite of the relatively large unit entropies. Figure 13 plots $M_{\text{normed}}$ for unitSizes 8, 12, and 16.

Two features evident in Figures 12 and 13 are the increase of Mnormed and the “drop-out” (to zero) of Mnormed with increasing unitSize. To understand the origin of these effects, note that the number of possible states, $S$, scales exponentially with unitSize $N$. Here, with four possible Ramachandran rotamers per residue, $S \approx 4^N$. Thus, as $N$ increases, the probability of all samples of the ensemble for a set of $N$ residues falling into the same state approaches unity, and then the entropy and all its properties are zero. This effect leads to the drop out feature in Figure 13. There are more half-units with zero entropy as the unitSize increases. Figure 12 shows that for some residue ranges, $H_{\text{ind}}$ is essentially half the value of other ranges, indicating that one of the half-units entropy is zero. The number of such half-units increases with $N$. These zero MI regions mirror the independent entropy plots in Figure 11.

The normed MI never becomes unity, indicating that on any scale of the unitSizes investigated, the sampling of units never becomes completely correlated; but the Mnormed values that are not zero do increase with unitSize. That reflects the definition of MI as the difference between independent and dependent entropies, both of which increase with increasing unitSize, and because the half-unit entropies are not increasing as fast as this difference, the normed MI will tend to increase.

The above analysis shows that the analyzed ensemble is indeed highly disordered. As long as there are units with sufficient sampling to give large entropy values, their MI tends to be very small. Thus, at this unit scale, $\alpha$Syn can be classified as essentially not-MoRF. As larger unitSizes are considered, when both half-units do still have entropy, then the MI values show that they are MoRFs. Those residue ranges for which the entropy falls to zero are then not-IDP. These latter residue ranges are around 20 in the N-terminal, 70 in the aggregation-prone, and 120 in the C-terminal regions. That the entropy in the aggregation prone region is much reduced may be related to its oligomerization capability by some specific geometric requirement. The entropy reduction around residue 20 may be associated with the amphipathic helical character of the N-terminal residues 1−25.$^{53}$

4. DISCUSSION

It is abundantly clear that IDPs exhibit a residual order versus random coil behavior.$^{61,73,75}$ In a given sequence, there may
be varying degrees of order in the subsequences that compose the sequence. The definition of conformations used here relies on residues sampling at least two distinct rotamers in the Ram(a) space and building a 2N residue subsequence conformational space from the possible Ram(a) rotamers. The MoRF definition in our work encapsulates the concept of a subsequence that has a large entropy via residues sampling distinct, rotameric states, yet because of the dependence among the sampled rotamers of the two halves of a given unit, there is substantial MI. The MI definition developed here is based on the statistical dependence between the two half-units that comprise the unit in Ram(2N) space—the 4N dimensional Φ/Ψ space for a subsequence of 2N residues. It is amenable to normalizing the extent of MI between different units and also between units of different unitSizes.

Focusing on ensemble data and analyzing units of varying sizes provide a structural view of IDPs that can be related to function. In particular, the conformational selection mechanism whereby a ligand selects from a limited number of distinct protein conformations can be explored by these methods. Different conformers of the MoRFs may be suitable for different ligands. The limited number of conformers sampled, as measured by MI, lends a specificity to the potential binding of different ligands.

Obtaining a reliable conformational ensemble is a difficult task, but such ensembles do permit the testing of structure—function hypotheses, much like X-ray structural data on enzymes lead to testable reaction mechanism hypotheses. The combination of experimental and simulation methods have provided some ensembles, and here we have analyzed ensemble data for Ub, Sic1, and αSyn. For Ub, a Ramachandran analysis [Ram(1)] of the residues found that, of the four ensembles, two, 2KOX and UB_2LJ5, did not have sufficient rotamer sampling to warrant our approach. Of the two that did, UB_ERIDU and UB_2K39, the UB_ERIDU ensemble was generated under denaturing conditions. That is very useful to validate our approach. Intuitively, a denatured ensemble should have a large entropy but essentially no MI for all unit sizes, as is found here.

The UB_2K39 ensemble data does, in contrast, show that there are MoRF units present for unitSizes 2–10. For unitSize 2, Table S1 of the Supporting Information shows 16 units with MInormed > 0.4 and, of these, there are 8 that are considered as MoRFs as the entropy Hind > 0.4 is relatively large. There are not-MoRF units, too, that is, with sufficient entropy but small MI. For these units, there is rotamer sampling of the residues but the sampling is relatively independent. Table 1 collects some of the residues that are MoRF and not-MoRF. The secondary structure designations given there do indicate that both MoRFs and not-MoRFs are concentrated in turn, bend, β-strand, and unassigned regions of the sequence. As the unitSize increases, one might anticipate a dropoff in the number of MoRFs, but at least to unitSize 10, the dropoff is mild. For unitSize 10, there are 4 MoRFs relative to 9 for the smaller unitSizes, whereas correspondingly, the number of not-MoRFs increases with increasing unitSize. For all unitSizes, the MoRF and not-MoRF units are concentrated in the residue ranges 0–10 (β-strand and turn) and 32–60 (turns, β-strand, and no secondary structure assigned). The tail region (no secondary structure) is strongly not-MoRF. These results are in accordance with the DSSP secondary structure assignment of Ub shown in Figure 3 as well as with the NMR RMSD data.

The Ub tail (residues 72–75) result of small MI is in good agreement with the Caliandro et al. analysis, based on long (0.2 μs) MD simulations, where they find large RMSDs of the tail residues. They note that these residues are involved in molecular recognition. Figure 4 shows that the Lange et al. UB_2K39 ensemble also implies large conformational fluctuations, as the entropy is the largest here. Our cluster results do a good job of separating the UB_2K39 ensemble into rather distinct conformations, emphasizing that clustering based on (1) backbone dihedrals and (2) using the product of individual Ram(a) rotamers to cluster the Ram(a) conformations is an effective strategy to produce sharp clusters. Note that the Ram(a) space is already of quite high dimension (dimension = 16), arising from Ram(a) Φ/Ψ and, for periodic dihedrals, clustering done in Fisher space, as discussed in Section 2 Methods.

As a general conclusion, the analysis does show that Ub exhibits extensive regions where residue sampling of rotameric states is robust, yet from the dependence between the half-units, the sampling is quite restricted. These results agree with and amplify previous studies of Ub. The Lange et al. (UB_2K39 ensemble) experimental design permitted observations on the microsecond timescale that, as they show, led to a more extensive exploration of configuration space reflected in quite a few residues sampling multiple rotamers. They demonstrated that the heterogeneity in their free Ub ensemble spans conformers that occur in a number of Ub-ligand conformers obtained by crystallography. As they note, it strongly suggests that a conformational selection hypothesis is operative. Candotti et al. in their urea denatured ensemble (UB_ERIDU) find an ensemble best characterized as random-coil-like. To reach this conclusion, they used a statistical coil model of an unfolded protein with their measured RDCs as restraints. That certainly will emphasize that the random coil nature and our zero MI agree well with their conclusions.

Two other ensembles, for the CDK inhibitor Sic1 and for the amyloid protein αSyn, were analyzed. Both are more IDP-like than the UB_2K39 Ub ensemble. For Sic1, the NMR and computationally derived ensemble display a wide range of conformational geometries but was shown to have significant nonrandom character. The MoRF analysis shows that the sequence has a large entropy and substantial MI values, thus supporting this conclusion. Contrasting the entropy and MI values around the phosphorylation sites and away from them suggest that the phosphorylation sites are more structured, as might be required for interaction with its receptor Cdc4. Comparing the scale of the entropy of this Sic1 ensemble with that of Ub shows that it is larger, indicating that, from this perspective, Sic1 is more IDP-like than Ub. The αSyn ensemble presents another case where experiment has shown that the IDP character is very strong but was argued to not be a random coil based on smaller radius of gyration and enhanced PPII sampling in the Ramachandran plane. The ensemble analyzed here does show enhanced PPII and β sheet propensities but does not reveal significant differences in them as a function of the sequence. Nevertheless, the feature that the aggregation-prone region (residues 60–95) has enhanced PPII and β sheet sampling is present. There is a strong effect of lack of entropy in some regions that is reflected in the not-MoRF character. This does occur in the N-terminal and the aggregation-prone regions, the latter region a source of oligomerization capability.
Some cautions are worth noting. First, our analysis is predicated on conformational entropy arising from distinct rotamer sampling in the residue Rama(1) space. This is distinct from entropic contributions from single rotamer populations that arise from the width of those distributions. There is an entropic effect from those sources, too. However, while entropy can be, with care, defined for continuous random variables, it does not lend itself to concepts such as MI. Focusing on different rotamer samplings, though, is well-suited to explore the role of conformational sampling in proteins and amenable to a MI approach. Second, the ensembles constructed using experimental data along with computational restraints tend to be rather small; they are minimal ensembles designed to fit the constraints. Thus, for example, the Ub_2K39 ensemble has 116, the Sic1 43, and the αSyn 576 members. Clearly, it is risky to make more than qualitative conclusions using a statistical approach from small ensembles. In particular, as the number of residues in the units considered increases, the exponential increase in the number of possible states shows that it is likely to have only one populated state. Then, all entropic properties are zero. Third, the division between MoRF, not-MoRF and not-IDP that we introduce depends on their definitions that are based on the values of unit entropy and MI. For example, for Ub unitSize 10 in the 41–54 region that is β-strand-turn/β-strand, there is, with reference to Table 1, actually a continuous transition between more MoRF to more not-MoRF behavior. With a larger ensemble and correspondingly better statistics, the entropic properties of the units may become smoother.

Finally, we speculate about whether some methods may lead to results that suppress the presence of MoRFS, that is, experiments and analysis methods where the obtained ensemble (1) does not display rotameric sampling, leading to negligible entropy or (2) while sampling rotamers miss the presence of MI. The methodology for the UB_2K39 ensemble analysis uses CONCOORD that creates an initial ensemble using geometric restraints based on structural data, and this ensemble is refined by the long-time experimental data. Here, one might anticipate that rotamers are well-sampled and dependencies embodied in MoRFS are preserved. By contrast, methods that start from a given X-ray structure and use simulated annealing (e.g., the UB_2KOX ensemble) may not be able to surmount barriers and result in low entropy, sampling essentially one rotamer in Rama(1) space. Other methods use (very) high-temperature MD to generate initial conformers that are then experimentally constrained. Here, while there may be a large conformational entropy from the high-temperature sampling, the dependencies may be suppressed in spite of the imposition of experimental restraints.

5. COMPUTATIONAL METHODS

5.1. Ub Data. Four NMR-based Ub ensembles were analyzed for suitability for this study. These are referred to as UB_2K39, UB_2KOX, UB_2LJS, and UB_ERIDU. Details of how they were obtained are given in the referenced articles. Briefly, UB_2K39 consists of 116 conformations obtained by refinement of RDCs and NOEs of free Ub that spans up to microsecond dynamics. UB_2KOX (UB_2LJS) consist of 640 (160) conformations obtained by ENSEMBLE restrained by NOEs and RDCs. The UB_ERIDU ensemble consists of 1000 structures based on NMR RDCs. It was generated by the ERIDU procedure that uses MD with a restraint penalty term that incorporates a distance from the experimental RDC data to generate a data-restrained ensemble. It was carried out on urea-denatured Ub.

For our purposes, what is needed is for some substantial number of residues to sample distinct rotameric states in the ensemble; for example, sampling the α helical and PPII regions. For each of these four ensembles, we first made Ramachandran “dot” plots of each residue (Ub has 76 residues; for Ramachandran plot purposes, the first and last residues are excluded). All four Ub do show population mainly in the PPII, β, α, αL, and what we will designate as O regions. Examination of these Ramachandran plots for the UB_2KOX and UB_2LJS ensembles only shows a few residues with rotameric sampling in the Ramachandran space to the degree that would make our methods appropriate. That is, for almost all residues, the Ramachandran space occupation of the ensemble falls in only one of the above population designations. Therefore, we will focus on the UB_ERIDU and UB_2K39 ensembles that exhibit multiple occupancies for most of the residues.

5.2. Sic1 Data. An ensemble for Sic1 consisting of 43 conformers was obtained from NMR data. Experiments monitored chemical shifts and 15N relaxation rates, RDCs, and PRE distance restraints as well as SAXS spectra. These data were combined with ENSEMBLE calculations to provide the conformational ensemble.

5.3. αSyn Data. An αSyn ensemble with 576 members was obtained from the Protein Ensemble Database, entry PED9AAC. PRE NMR measurements were carried out with an inserted cysteine residue. The measured distance restraints were incorporated as restraints in a high-temperature MDs simulation to generate the conformers.

5.4. MoRF Definition from Entropy and MI. We will refer to a subsequence formed from a contiguous set of 2N (N = 1, 2, ...) residues as a unit of unitSize 2N.

Definitions:

MoRF: A unit that has substantial total entropy and a strong dependence (large MI) among the conformations that these residues sample.

Not-MoRF: A unit with a large entropy but a small MI value, thus sampling rotameric states but in an essentially independent fashion.

Not-IDP: A unit with a small entropy and a small or large MI value. Essentially one conformation is sampled.

To provide a well-posed definition of MI and entropy, the state space will be discretized. Conformational states are then defined as the direct product of Ramachandran-discretized rotamer populations. For notation, we define Rama(1) as the rotamer sampling of one residue in Ramachandran (ϕ/ψ) dihedral space and Rama(N) for the rotameric sampling of N contiguous residues. For example, dividing Rama(1) space into four discrete regions produces, for N contiguous residues, a Rama(N) state space of dimension 4N.

To implement the MoRF definition, let X denote an N-dimensional random variable and x = (X1, X2, ..., XN) ∈ Ω(X) a member of the set of possible (discrete) values {X} that it can take on. The corresponding (information) entropy is

$$H(X) = - \sum_{x \in \Omega(X)} p(x) \ln p(x)$$

For random variables X and Y in the 2N dimensional Ramachandran space, an information-entropy-based MI MI(X;Y) can be defined as
The mutual information (MI) is defined between random variables $X$ and $Y$ as

$$MI(X; Y) = \sum_{x \in \{X\}} \sum_{y \in \{Y\}} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)}$$  \hspace{1cm} (5.2)

This is a generalization of the usual MI definition\textsuperscript{50} to random variables that are composite, desired here, as appropriate to a definition of a MoRF. Note that MI is a specialization of the Kullback–Liebler entropy\textsuperscript{77} because the denominator probability is an independent probability versus some other arbitrary probability $q(x,y)$. By definition

$$MI(X; Y) = H(X) + H(Y) - H(X, Y)$$

$$\equiv H_{\text{ind}} - H_{\text{dep}}$$  \hspace{1cm} (5.3)

thus MI is the difference between independent and dependent entropies.

MI has the virtue (vs a correlation coefficient) of quantitating the amount of dependence of one random variable on another random variable. The drawback is that when comparing MI for different sets of variables, the varying entropy contributing to the different MI makes comparison difficult. However, it is straightforward to produce upper and lower bounds on each MI. Clearly, a lower bound is zero, achieved when the two random variables are independent. An upper bound can be formulated\textsuperscript{59}

$$MI(X; Y) \leq \min(H(X), H(Y))$$  \hspace{1cm} (5.4)

with the special case that if $\min(H(X), H(Y)) = 0$, then $MI(X; Y) = 0$. The maximal entropy reduction from complete dependence is the smaller of the two entropies. For example, if $H(Y) < H(X)$ and there is the functional relation $Y = f(X)$, then $MI(X; Y) = H(Y)$ and $H(X, Y) = H(X)$. No new statistical information is available when there is a functional relation between $Y$ and $X$. In general, it is worth noting that because $MI(X; Y) \geq 0$, it follows from $MI(X; Y) = H(X) - H(X|Y)$, with $H(X|Y)$ being the conditional entropy, that conditioning (sensibly) reduces entropy; $H(X|Y) \leq H(X)$.

To put the MI of different MoRFs on the same scale it then is useful to define

$$MI_{\text{normed}}(X; Y) = MI(X; Y)/\min((H(X), H(Y)))$$  \hspace{1cm} (5.5)

that satisfies

$$0 \leq MI_{\text{normed}}(X; Y) \leq 1$$  \hspace{1cm} (5.6)

Then, comparing the degrees of dependence between different MoRFs is straightforward.

The MI definition in eq 5.2 is different from the conventional one, where the comparison is made with the independence of all random variables. The conventional MI expression denoted as $MI_{\text{con}}(X,Y)$ is

$$MI_{\text{con}}(X; Y) = \sum_{x \in \{X\}} \sum_{y \in \{Y\}} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)}$$  \hspace{1cm} (5.7)

with $p_{\text{ind}}(x,y)$ being the product of the probabilities of the 2N random variables. In the Supporting Information, we show that $MI(X; Y)$ can be expressed in terms of $MI_{\text{con}}(X,Y)$ for $X$ and $Y$ ranging from 1 to $N$ random variables. For MoRF comparison purposes, $MI_{\text{con}}(X,Y)$ is not as useful as the MI of eq 5.2 because normalizing it is not so straightforward.

5.5. Clustering Algorithm for MoRFs. Given the ensemble data of a sequence’s backbone $\varphi$ and $\psi$ dihedral angles, a classification into discrete rotamers for each residue is required. This classification is carried out in the Rama(1) space—the $\varphi/\psi$ dihedrals of each residue. In this analysis, we will lump the PPII and $\beta$ into one rotameric state, as, for a given residue, it is hard to distinguish those two regions in the Ub ensemble data. Thus, four regions will be used: they are designated as and centered on $\alpha$ (−90,−30), $\beta$-PPII (−120,150), aL (90,45), and O (90,120).

Some clustering algorithm must be used to obtain the MoRFs in the 2N dimensional Rama(2N) space. Previously, a “compositional clustering” method was introduced whereby each dihedral angle of interest is separately clustered into its discrete rotamers and a state of a sequence was defined as the direct product of each rotamer’s conformation\textsuperscript{8} (the clustering of each dihedral angle is carried out in “Fisher” coordinates\textsuperscript{79} that represent each angle in terms of its sine and cosine to provide a proper metric in the periodic dihedral space). In this discrete basis, a state characterizes the conformation of the sequence for each trajectory sample. The state populations are then readily obtained from their proportions in the trajectory. The virtue of compositional clustering in the dihedral space is that each dihedral is separately accounted for versus typical global clustering algorithms that cluster by summing over a set of coordinates. Here, the clustering of each residue is done in Rama(1). The states of 2N dimensional units can be found, along with the N-dimensional states of the units’ first and last halves, forming the putative MoRF, using the compositional clustering method. The MI can be evaluated from eq 5.2 using the samples of these three sets of states, along with the entropies $H(X)$ and $H(Y)$, to obtain $MI_{\text{normed}}(X,Y)$ of eq 5.6.

5.6. Protocol for Obtaining MoRFs. All data used in this Ub study were obtained as pdb files, as described in Section 2.1. The Biopython\textsuperscript{80} or Analyzer\textsuperscript{81} software was used to generate the dihedral backbone angle samples from the pdb files. The Rama(1) samples of each residue were concatenated into Rama(2N) data files and used as input to our k means program that does the compositional clustering in the Fisher representation of each residue. For a unit of unitSize 2N, the clusters are in a $2(\varphi/\psi) \times 2(\text{Fisher}) \times 2N(\text{unitSize dimension}) = 8N$ dimensional space. We examined unitSizes 2N = 2, 4, 6, 8, and 10 so that the maximum cluster dimension is 40. If the clustering were done directly on the units, as some experimentation showed, the sharpness of clustering arising from the rotameric sampling in Rama(1) space would be greatly reduced.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b00898.

Derivation of the relation between the two definitions of MI and tables of MoRF/not-MoRF for unitSizes 2, 4, 8, and 10 (PDF)

AUTHOR INFORMATION

*Corresponding Author

E-mail: cukier@chemistry.msu.edu. Phone: 517-353-1170. Fax: 517-353-1793.

ORCID

Robert I. Cukier: 0000-0003-0306-7246
Notes
The author declares no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge the Department of Chemistry and High Performance Computing Center facilities at the Michigan State University.

REFERENCES

(1) Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. Nat. Rev. Mol. Cell Biol. 2005, 6, 197–208.
(2) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing Protein Intrinsic Disorder. Chem. Rev. 2014, 114, 6561–6588.
(3) Uversky, V. N. Unusual Biophysics of Intrinsically Disordered Proteins. BBA. Biochem. Biophys. Acta. Proteins Proteomics 2013, 1834, 932–951.
(4) Wright, P. E.; Dyson, H. J. Linking Folding and Binding. Curr. Opin. Struct. Biol. 2009, 19, 31–38.
(5) Nussinov, R.; Tsai, C.-J.; Ma, B. The Underappreciated Role of Allostery in the Cellular Network. Annu. Rev. Biophys. 2013, 42, 169–189.
(6) Mao, A. H.; Lyle, N.; Pappu, R. V. Describing Sequence-Ensemble Relationships for Intrinsically Disordered Proteins. Biochem. J. 2013, 449, 307–318.
(7) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. J. Mol. Biol. 1999, 293, 321–331.
(8) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradović, Z. Intrinsic Disorder and Protein Function. Biochemistry 2002, 41, 6573–6582.
(9) Fink, A. L. Natively Unfolded Proteins. Curr. Opin. Struct. Biol. 2005, 15, 35–41.
(10) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically Disordered Proteins: Regulation and Disease. Curr. Opin. Struct. Biol. 2011, 21, 432–440.
(11) Glad, A. M.; Stanglciu, L.; Radulescu, A.; Holder, O.; Monkenbusch, M.; Biehl, R.; Richter, D. Internal Nanosecond Dynamics in the Intrinsically Disordered Myelin Basic Protein. J. Am. Chem. Soc. 2014, 136, 6987–6994.
(12) Csermely, P.; Palotai, R.; Nussinov, R. Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. Trends Biochem. Sci. 2010, 35, 539–546.
(13) Cozzetto, D.; Jones, D. T. The Contribution of Intrinsic Disorder Prediction to the Elucidation of Protein Function. Curr. Opin. Struct. Biol. 2013, 23, 467–472.
(14) Drobnak, I.; De Jonge, N.; Haesaerts, S.; Veninauer, G.; Loris, R.; Lah, J. Energetic Basis of UNCoupling Folding from Binding for an Intrinsically Disordered Protein. J. Am. Chem. Soc. 2013, 135, 1288–1294.
(15) Huang, Y.; Liu, Z. Kinetic Advantage of Intrinsically Disordered Proteins in Coupled Folding-Binding Process: A Critical Assessment of the “Fly-Casting” Mechanism. J. Mol. Biol. 2009, 393, 1143–1159.
(16) Rogers, J. M.; Steward, A.; Clarke, J. Folding and Binding of an Intrinsically Disordered Protein: Fast, but Not “Diffusion-Limited”. J. Am. Chem. Soc. 2013, 135, 1415–1422.
(17) Ubbink, M. The Courtship of Proteins: Understanding the Encounter Complex. FEBS Lett. 2009, 583, 1060–1066.
(18) Chen, J. Towards the Physical Basis of How Intrinsic Disorder Mediates Protein Function. Arch. Biochem. Biophys. 2012, 524, 123–131.
(19) Davey, N. E.; Van Roey, K.; Weatheritt, R. J.; Toedt, G.; Uyar, B.; Ahlenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T. J. Attributes of Short Linear Motifs. Mol. Biol. Syst. 2012, 8, 268–281.
(20) Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annu. Rev. Biochem. 2014, 83, 553–584.
(21) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. 2014, 114, 6589–6631.
(22) Asgari, E.; Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS One 2015, 10, No. e0141287.
(23) Deng, X.; Eckoldt, J.; Cheng, J. A Comprehensive Overview of Computational Protein Disorder Prediction Methods. Mol. BioSyst. 2012, 8, 114–121.
(24) Das, R. K.; Pappu, R. V. Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues. Proc. Natl. Acad. Sci. U.S.A. 2013, 110, 13392–13397.
(25) Fisher, C. K.; Stultz, C. M. Constructing Ensembles for Intrinsically Disordered Proteins. Curr. Opin. Struct. Biol. 2011, 21, 426–431.
(26) Tompa, P. Unstructured Biology Coming of Age. Curr. Opin. Struct. Biol. 2011, 21, 419–425.
(27) Ozsen, V.; Bauer, F.; Salman, L.; Huang, J.-r.; Jensen, M. R.; Segard, S.; Bernado, P.; Charavy, C.; Blackledge, M. Flexible-Meccano: A Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and Their Associated Experimental Observables. Bioinformatics 2012, 28, 1463–1470.
(28) Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-Ray Scattering. J. Am. Chem. Soc. 2007, 129, 5656–5664.
(29) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. A Structural Model for Unfolded Proteins from Residual Dipolar Couplings and Small-Angle X-Ray Scattering. Proc. Natl. Acad. Sci. U.S.A. 2005, 102, 17002–17007.
(30) Montalvao, R. W.; De Simone, A.; Vendruscolo, M. Determination of Structural Fluctuations of Proteins from Structure-Based Calculations of Residual Dipolar Couplings. J. Biomol. NMR 2012, 53, 281–292.
(31) Ángyán, A. F.; Gáspári, Z. Ensemble-Based Interpretations of NMR Structural Data to Describe Protein Internal Dynamics. Molecules 2013, 18, 10548–10567.
(32) de Groot, B. L.; van Aalten, D. M. F.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. C. Prediction of Protein Conformational Freedom from Distance Constraints. Proteins: Struct., Funct., Genet. 1997, 29, 240–251.
(33) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts. J. Am. Chem. Soc. 2010, 132, 1270–1272.
(34) Krzeminski, M.; Marsh, J. A.; Neale, C.; Choy, W.-Y.; Forman-Kay, J. D. Characterization of Disordered Proteins with Ensemble. Bioinformatics 2013, 29, 398–399.
(35) Allison, J. R. Using Simulation to Interpret Experimental Data in Terms of Protein Conformational Ensembles. Curr. Opin. Struct. Biol. 2017, 43, 79–87.
(36) Levine, Z. A.; Shea, J.-E. Simulations of Disordered Proteins and Systems with Conformational Heterogeneity. Curr. Opin. Struct. Biol. 2017, 43, 95–103.
(37) Frenkel, D.; Smit, B. Understanding Molecular Simulation: From Algorithms to Applications: Academic Press: San Diego, 1996; pp 1–638.
(38) McCammon, J. A.; Harvey, S. C. Dynamics of Proteins and Nucleic Acids; Cambridge University Press: Cambridge, 1987; pp 1–234.
(39) Bax, A.; Best, R. B. Computational and Theoretical Advances in Studies of Intrinsically Disordered Proteins. Curr. Opin. Struct. Biol. 2017, 42, 147–154.
(40) Henriques, J.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: On the Accuracy of the Tip4p-D Water Model and the Representativeness of Protein Disorder Models. J. Chem. Theory Comput. 2016, 12, 3407–3415.
(81) Lou, H.; Cukier, R. I. Analyzer 2.0, East Lansing, 2008.