Review

Israel Journal of Chemistry

DOI: 10.1002/ijch.201400030

Structural Heterogeneity and Dynamics of the Unfolded Ensemble

Ignacia Echeverria*[a] and Garegin A. Papoian*[a, b]

Abstract: Significant efforts have been devoted to understanding the structural and physicochemical properties of unfolded and intrinsically disordered proteins. Combining experimental measurements with molecular simulations and polymer theory calculations has emerged as a powerful route to accurately characterize the rapidly interchanging conformations of the unfolded ensemble. We review a selection of recent works on the dynamics of unfolded and intrinsically disordered proteins, focusing primarily on computer simulation and theoretical approaches. We use the energy landscapes paradigm to highlight various computational techniques and to outline several directions for future research. One major, immediate challenge is to gain deeper insights into the nature of the energy barriers that determine the roughness of the energy landscape of unfolded proteins. A second important challenge is to better characterize and understand the functional role of partial ordering, or alternatively, disorder-to-disorder transitions, between various phases of the unfolded state.

Keywords: computational chemistry · energy landscapes · protein folding · proteins · unfolded proteins

1. Introduction

Better understanding of the structural and dynamical properties of the unfolded state will provide a more complete picture of the way proteins fold. Likewise, gaining insights into the unfolded state is essential to understanding the properties and functions of the proteins that are known to be disordered in their biologically active form. However, atomically detailed information of unfolded proteins remains sparse, as a result of the conformational heterogeneity and dynamics of the unfolded ensemble. Several recent works have strived to address this gap.

The initial interest in the unfolded state was largely motivated by the desire to better understand the thermodynamic driving forces governing protein folding. For most small globular proteins at equilibrium, only two populations, unfolded (U) and folded (F), predominate, with only negligible thermodynamic contributions of the intermediate states. In this case, the folding reaction can be written as $U \rightleftharpoons N$, with an equilibrium constant of $K_{eq} = [N]/[U]$. This, from a physical perspective, is equivalent to considering that there exist two equally stable phases of the protein chain which are separated by a free energy barrier and consequently, the folded phase cannot decay; for example, via continuous swelling.[1] However, it is currently thought that the early stages of protein folding kinetics are dominated by the structural collapse into globular or molten globule conformations, followed by two-state folding kinetics.[2–4]

Globular proteins and some intrinsically disordered proteins (IDPs), under appropriate temperature and solvent conditions, undergo a collapse transition into the molten globule state. This transition leads to an ensemble of collapsed structures, whose average size depends continuously on, for example, the denaturant concentration.[5] This result suggests that, for flexible heteropolymers, such as proteins, coil-to-globule transitions are smooth second-order phase transitions.[6] Proteins in the molten globule state are rather dense, with native-like secondary structural elements, and are characterized by the absence of side-chain close packing, where instead, the side-chains dynamically explore various rotameric states. The key thermodynamic features of this state are the favorable increase in conformational entropy compared with the folded ensemble, at the expense of less favorable intramolecular interactions.[7–9] The collapse into the molten globule state, which is mediated mostly by hydrophobic interactions, is commonly much faster than the overall folding time.[10] Once a protein has collapsed to an unstructured globule, local motion will drive the conforma-

[a] I. Echeverria, G. A. Papoian
Department of Chemistry and Biochemistry
University of Maryland
College Park, MD 20742 (USA)
e-mail: iechever@umd.edu

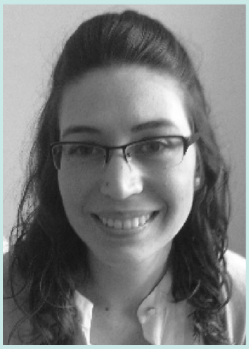
[b] G. A. Papoian
Institute for Physical Science and Technology
University of Maryland
College Park, MD 20742 (USA)
e-mail: gpapoian@umd.edu

tional search towards the native structural basin.[10] The timescales over which unfolded proteins undergo substantial conformational reconfiguration, which is limited by the diffusive encounter of the parts of the polypeptide chain, ultimately determine the "speed limit", or the maximum rate, at which proteins can potentially fold.[11,12] Consequently, the upper limit for the kinetics of a protein folding reaction will be governed by the rate at which intra-chain contacts are formed and broken in the molten globule state.

According to the Kramers' theory, the barrier crossing rate, k, for unimolecular reactions in a damped environment varies as $k \propto \gamma^{-1} \exp(-\triangle G/k_B T)$, where $\triangle G$ is the barrier height and γ is the friction coefficient. Here, the pre-exponential factor sets the "speed limit" of the folding reaction, which has been estimated to be of the order of 1 μs.[11–13] This estimate is comparable to the measured folding times of the fastest folding proteins,[11,12] suggesting that the structural properties and dynamics of the unfolded state can affect the folding rates.[14] Moreover, over the last few years, the dynamics of unfolded proteins have been characterized under different denaturant and solvent viscosity conditions, allowing researchers to single out the different contributors to the reconfiguration dynamics in the unfolded state. These studies have suggested that internal friction, which correlates with the compactness of the unfolded protein, rather than solvent friction, may provide the dominant contribution in the folding kinetics of proteins that fold in the microsecond range or faster.[15,16]

A renewed interest in understating the properties of unfolded proteins has come with the realization that between one-sixth and one-third of eukaryotic proteins are either disordered or have large disordered regions.[17,18] Moreover, these proteins have been identified as key players in a variety of biological processes, such as transcription and signal transduction,[19] and their misregulation has been linked to numerous diseases. From a structural perspective, disordered proteins populate an ensemble of heterogeneous conformations, which is determined by a balance of the inter-residue interactions and the entropy of the chain. It has been suggested that the net charge per residue modulates the ensemble of populated conformations, which can range from extended to highly collapsed in aqueous solvents.[20]

Among the properties of the unfolded state that remain to be fully understood are the roles of conformational heterogeneity and residual structures in the thermodynamics and kinetics of folding and aggregation. For example, to what degree do the conformational heterogeneity and the residual secondary structure of the unfolded state govern the disorder-to-order transitions in globular proteins and in IDPs? Additionally, what is the origin and what are the properties of the energy barriers that determine the roughness of the energy landscape of unfolded and disordered proteins? This review focuses on the recent developments in characterizing the unfolded protein ensemble using molecular dynamics (MD) simulations and other computational methods. Additionally, we will discuss how computational methods can be used side-by-side with recently developed experimental techniques to further our understanding of the structural heterogeneity and reconfiguration dynamics of the unfolded ensemble. We will also discuss the similarities and differences of the unfolded ensemble of globular proteins and intrinsically disordered proteins.



Ignacia Echeverria was born in Santiago, Chile in 1982. She received her doctorate in molecular biophysics at the Johns Hopkins University in 2011, working in the laboratory of Professor Mario Amzel. During this period, she worked on developing methods to compute free energy differences using non-equilibrium molecular dynamics simulations. After her Ph.D., she joined the group of Professor Garegin Papoian at the University of Maryland as a post-doctoral researcher. During this time, she has worked on understanding the reconfiguration dynamics of unfolded proteins, the molecular origins of internal friction, and on determining the thermodynamics and kinetics of protein-DNA interactions.

Garegin A. Papoian was born in Yerevan, Armenia in 1973. He completed 4 years of undergraduate studies at the Russian Academy of Sciences, followed by graduate work in quantum chemistry with Professor Roald Hoffmann at Cornell University. He received his Ph.D. in 1999. He continued with post-doctoral work with Dr. Michael Klein and Dr. Peter Wolynes, studying protein physics. He has held faculty positions at the University of North Carolina at Chapel Hill, where he was tenured in 2010, followed by being the first Monroe Martin Professor at the University of Maryland. His group has been developing physico-chemical models of cytoskeletal dynamics, chromatin assembly, and protein motion and function.

2. Characterizing the Structural Heterogeneity of the Unfolded Ensemble

A defining characteristic of unfolded proteins and IDPs is their conformational heterogeneity. The unfolded ensemble can be grossly characterized by its polymeric properties, such as average size, shape or density. For example, the radius of gyration (R_g), which provides a measure of a chain's global dimensions, is suitable to identify coil-to-globule transitions.[21–23] However, these measures only ac-

count for global and averaged properties, and do not capture many important aspects of the conformational heterogeneity of the unfolded ensemble. Hence, it is desirable to develop an order parameter that will, for example, identify local changes in the structure that do not necessarily change the global properties of the polypeptide and that, additionally, is capable of distinguishing between homogeneous and non-homogenous ensembles of conformations. A proper order parameter would allow the quantification of the degree of heterogeneity of ensembles simulated at different conditions, such as temperature and denaturant concentrations, or upon changing the peptide by mutations or post-translational modifications.

Several measures have been proposed to characterize the unfolded ensemble. For example, the pairwise-q, an often used order parameter in spin glass physics[24] and in protein folding,[25–27]

$$q_{ij} = \frac{1}{N_{pairs}} \sum_{a,b} \exp\left[-\frac{\left(r^{i}_{a,b} - r^{j}_{a,b} \right)}{2\sigma^2} \right] \quad (1)$$

quantifies the structural similarity between two conformations i and j. Here $r^{i}_{a,b}$ and $r^{j}_{a,b}$ are the pairwise distances between C_α atoms, a and b, in the conformations i and j, respectively; N_{pairs} is the total number of C_α pairs; and σ is a parameter that controls the resolution of the order parameter, usually set to 2 Å.[28] q_{ij} ranges from near zero, when two conformations have no structural resemblances, to 1, when two structures are identical. The structural heterogeneity of an ensemble of conformations can by quantified by creating a histogram of q for all pairs of conformations sampled, to produce the probability distribution, $P(q)$.[28] The shape and position of the distribution provides a fingerprint of the heterogeneity of the sampled ensemble. For example, comparing the $P(q)$ of the conformational ensembles of wild-type and acetylated H4 histone tail Potoyan and Papoian determined that acetylation induces a significant reorganization of the energy landscape, by inducing partial structural ordering.[28] In this particular example, the bi-modal shape of $P(q)$ signals that the ensemble contains distinct populations, with one of the populations (higher pairwise-q) being subject to structural ordering, and another being more disordered (lower pairwise-q).

Another related order parameter, developed by Lyle *et al.*,[22] was designed to discriminate between systems where the coil-to-globule transition is coupled to the folding reaction and those which undergo coil-to-globule transitions, adopting a highly heterogeneous ensemble of compact conformations. This order parameter quantifies the degree of similarity between all pairs of conformations in the ensemble by converting each conformation to a vector, whose components are the inter-residue distances $\mathbf{V}_C = \{r_{12}, r_{13}, \ldots, r_{N-1,N}\}$. Here, r_{ij} can be either the distance between the C_αs of residues i and j, or defined as:

$$d_{ij} = \frac{1}{Z_{ij}} \sum_{m \in i} \sum_{n \in j} \left| \mathbf{r}^{i}_{m} - \mathbf{r}^{j}_{n} \right| \quad (2)$$

Here, $\mathbf{r}^{i}_{m}$ and $\mathbf{r}^{j}_{n}$ are the position vectors of atoms m and n within residues i and j, respectively, and Z_{ij} is the number of unique pairs of interatomic distances between residues i and j. The conformations, a and b, are compared using the pairwise dissimilarity measure, defined by projecting the conformational vectors, $\mathbf{V}_C$:

$$D_{ab} = 1 - \cos(\Omega_{ab}) \quad (3)$$

where $\cos(\Omega_{ab}) = \frac{V_a \cdot V_b}{|V_a||V_b|}$. Using this order parameter, the probability distribution $P(D)$ can be used to characterize the conformational heterogeneity of the ensemble in an analogous way to the one described above for $P(q)$. Nevertheless, to compare the ensembles obtained under different conditions or for different polypeptides, the average $\langle D \rangle$ needs to be normalized with respect to the maximum heterogeneity possible. For this, the authors suggest using an approximation of the Flory random coil model that is constructed for each polypeptide sequence.

A different approach to quantifying the structural heterogeneity of the unfolded ensemble is to perform structural clustering, according to their mutual RMSD,[29] of the sampled conformations. Using long molecular dynamics simulations, where multiple folding events are observed, Deng *et al.* determined the heterogeneity of the structures sampled between two adjacent folding/refolding events using the R_g as the order parameter to characterize the structural reorganization among the different sampled clusters. However, this approach would not necessarily identify local changes in the structure.

3. Structure within the Unfolded Ensemble

The dimensions and conformational heterogeneity of the unfolded proteins and IDPs are highly dependent on the environmental and solvent conditions such as temperature, denaturant and salt concentrations, and on their amino acid composition. For example, the chemically-induced unfolded state may be quite different from the unfolded state under more physiologically-relevant solvent conditions. At near-zero and zero denaturant concentrations, proteins adopt collapsed structures, with a large number of inter-residue contacts, which gives rise to secondary structures that can resemble those of their native states.[30–32] In contrast, at high denaturant concentrations, very few secondary structure elements are present,[33] and the conformations of the unfolded proteins can be well described by random coil models. IDPs, having high net charge and a low content of hydrophobic residues,[18,20] can exhibit extended conformations under physiological conditions.[34] However, about one-fourth of IDPs adopt

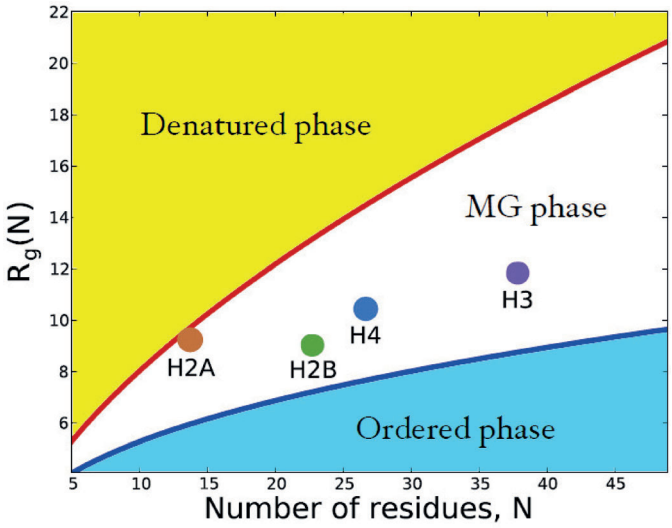


Figure 1. Phase diagram showing denatured, molten globular, and native globular regions for proteins. The horizontal axis is the number of residues of the polypeptide chain. The vertical axis is the average radius of gyration ($R_g(N)$). Disordered histone tails, which are indicated with the filled circles, populate relatively compact molten globule states. Reproduced with permission from Refs. [28, 53].

globule-like conformations without necessarily acquiring a homogenous ensemble of conformations (Figure 1).[35] It has been proposed, for certain IDPs, that the intrinsic heterogeneity of the sampled conformations is directly related to their function.[19,36]

Several studies of the unfolded ensemble have characterized the residual secondary structure, in an attempt to decipher the early stages of protein folding and protein stability.[31,37,38] Using molecular dynamics simulations, Tripathi *et al.*[39] investigated the unfolded ensemble of two variants of the all-beta sheet domain, Tenascin-C. This study demonstrated that early native contacts formed in the unfolded state, together with the local structural rearrangements, are fundamental to forming the folding nucleus. Additionally, differences in the residual secondary structure among the two variants causes backtracking, or local unfolding, in the early stages of folding, possibly slowing down folding kinetics.[39] The denatured state of the four-helix bundle Acyl-CoA binding protein, at very low denaturant concentration, has also been characterized as having residual secondary structure, a highly compact hydrophobic core, and an enriched number of long-range interactions.[40,41] Moreover, these studies have demonstrated that the unfolded ensemble is highly heterogeneous, exhibiting slow dynamics, and that a single mutation can perturb the folding core.[40] In another study, Maisuradze *et al.* studied the folding of the B-domain of the staphylococcal protein at different temperatures, finding that the conformational ensemble of the unfolded protein also contains a collection of conformations with residual secondary structure elements and native-like clusters of non-polar residues.[42]

Some IDPs, under specific conditions, upon association with a specific partner or by self-assembly, can adopt well-defined three-dimensional structures by coupling binding and folding.[36,43,44] It has been proposed that, for IDPs, the ability to adopt functional forms upon binding is determined by their conformational heterogeneities.[36] This, in the context of the energy landscape theory, is achieved by having a weakly funneled energy landscape that stabilizes the folded conformation only when additional interactions are present; for example, via molecular recognition of a specific interface.[45] Currently, it is not fully understood whether binding induces the IDP's structural changes or if binding selects the conformer with the "appropriate" structure. This is analogous to the "induced fit" versus "conformational selection"controversy in protein allostery, which has a long history.[46] On the other hand, some IDPs do not undergo disorder-to-order transitions when functionally active. Some examples of proteins that maintain various degrees of disorder in the bound state have been described in the last few years,[47] shedding light on the possibility of having a continuous structural spectrum of functional proteins.[45,48]

Molecular dynamics simulations have proven to be a powerful tool for studying conformational heterogeneity of IDPs,[49,50] providing the spatial and temporal resolution necessary to understand, for example, the molecular mechanism by which these proteins couple folding to binding in a functional way.[36,51,52] Histone tails, which play a crucial role in DNA compaction, are an example of IDPs which exhibit mostly compact molten-globule like structures (Figure 1). By studying the energy landscape of these proteins, Potoyan and Papoian showed that their conformational dynamics are highly organized, giving rise to ensembles of conformations that are far from being completely disordered, and which contain organized secondary structure elements, with well-defined conformational basins of attraction.[53] Further analysis showed that a post-translational modification, a widely used mode of biological regulation, can significantly modify the conformational and binding propensities of the histone tails by remodeling their energy landscapes.[28,53]

Molecular dynamics simulations are typically a single molecule technique. In recent years, new efforts have tried to elucidate the role of crowding and confinement in folding and aggregation.[54–56] In a first approach, by simulating the unfolding of a small ensemble of a ultra-fast-folding and -unfolding protein (32 copies of the engrailed homeodomain), McCully *et al.*[57] determined that the presence of neighboring protein molecules does not significantly alter the unfolding pathway, but does affect its kinetics by slowing down the process. As new advances in computational techniques allow simulating larger systems for longer timescales, the effect of crowders and

neighboring proteins in defining the unfolded ensemble should be investigated in more depth.

4. Reconfiguration Dynamics

At various stages of the protein folding reaction, the types and timescales of polypeptide chain motions can vary dramatically (see the review in Ref. [58]). The thermodynamic properties of the unfolded ensemble are determined by the roughness of the free energy surface, whereas the dynamical properties of the polypeptide chain are described by diffusive motions on the same surface.[26,27,59–61] In the unfolded state, proteins explore non-funneled or weakly-funneled regions of the energy landscape,[45,62] using diffusive motions that involve crossing a variety of micro-barriers to structural reorganization (Figure 2), determined by the intramolecular interactions and the degree of compaction of the polypeptide. For example, the hydrophobic collapse of unfolded proteins into molten globule conformations slows the reconfiguration dynamics, due to an increased roughness of the free energy landscape arising from increased steric clashes, backbone dihedral transitions, hydrophobic interactions, and the making and breaking of hydrogen bonds, among others.[62] In contrast, some IDPs, which have an increased number of charged residues, may adopt expanded conformations, where the roughness of the energy landscape is determined by long-range interactions.[45] This suggests that different dynamical regimes should be expected for different unfolded ensembles.

Even though it has been well established that large-scale motions of unfolded proteins are necessarily diffusive, it is unclear how much these motions are controlled by viscous drag exerted by the solvent, as compared to the effect of the inherent intramolecular energy landscape or "internal friction".[63,64] Experimentally, the contribution of the internal friction to the reconfiguration dynamics of unfolded proteins has been addressed by measuring the viscosity dependence of various timescales of protein dynamics (e.g., the folding time or reconfiguration time in the unfolded state).[15,16,65–67] In such experiments, when the reconfiguration time, τ, is found to depend linearly on the solvent viscosity η, i.e.,

$$\tau = a\eta + \tau_i, \tag{4}$$

then the zero-viscosity intercept, τ_i, is usually attributed to the internal friction.[15,16] These experiments have also determined that the magnitude of internal friction correlates with the compactness of the unfolded protein, dominating the reconfiguration time of the compact states in low denaturant concentration, and becoming negligible at high denaturant concentrations and for intrinsically disordered proteins that have expanded conformations (Figure 3). Furthermore, Soranno *et al.* were able to quantify the relative internal friction contributions as a function of denaturant concentration using polymer physics modeling of the corresponding experimental measurements. Here, a modified version of the Rouse model, the compacted Rouse model with internal friction (CRIF) was used.[15,68] This model considers a polymer as a collection of $N+1$ beads with coordinates $\mathbf{r}=(\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_N)$. Adjacent beads are connected by harmonic springs, such that the potential energy is given by:

$$V_0 = \frac{1}{2}k_0 \sum_{n=0}^{N} (\mathbf{r}_i - \mathbf{r}_{i-1})^2 \tag{5}$$

where $\mathbf{r}$ is the vector whose components are the bead positions and k_0 is the spring constant. To account for the polymer collapse upon lowering the denaturant concen-

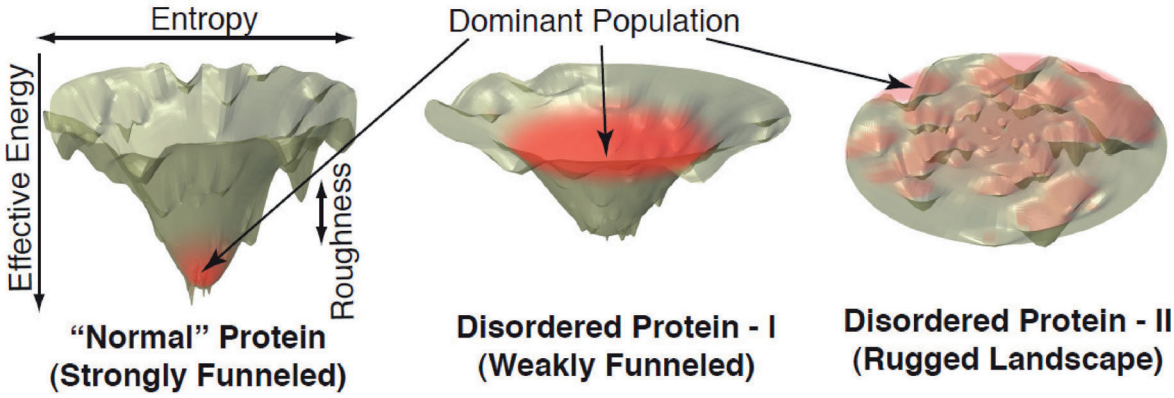


Figure 2. Free energy landscapes for globular and unfolded proteins. Left: Energy landscapes of globular proteins are thought to be funneled, such that the native state is both a thermodynamic global minimum and is also kinetically accessible[26,27]. Center: Energy landscapes of certain IDPs can be weakly funneled; for example, IDPs that upon binding, post-translational modification or aggregation, can adopt well-defined three-dimensional structures. Right: A random energy landscape where no specific functional conformation is favored. The dominant populations, or the expected ensemble of populated conformations is shown in red. Reproduced with permission from Ref. [45].

tration, the model includes a central potential which effectively compresses the chain towards the coordinate origin:

$$V_c = \frac{1}{2} k_c \sum_{n=0}^{N} r_i^2 \tag{6}$$

where the spring constant, k_c, controls the degree of the chain collapse. The internal friction is included by adding a drag force term that opposes the relative motion of two neighboring beads (i.e., $-d(\mathbf{r}_i - \mathbf{r}_{i-1})/dt$). Hence, the total force on the i^{th} bead, due to internal friction, is proportional to $-d(2\mathbf{r}_{i-1} - 2\mathbf{r}_i + \mathbf{r}_{i+1})/dt$. The dynamics of the CRIF chain is governed by:

$$-\xi_S \frac{d\mathbf{r}}{dt} - \xi_I \mathbf{k} \frac{d\mathbf{r}}{dt} - k_0 \mathbf{k}\mathbf{r} - k_0 \mathbf{I}\mathbf{r} + \mathbf{f}(t) = 0. \tag{7}$$

Here, ξ_S and ξ_I are, respectively, the solvent and internal friction coefficients, $\mathbf{f}(t)$ is a random force vector satisfying the appropriate fluctuation-dissipation relationship, $\mathbf{I}$ is the identity matrix and $\mathbf{k}$ is the connectivity matrix.[15,68]

Although several studies have characterized the presence of internal friction in proteins and polymers, the molecular origins of internal friction are currently not well understood. Recent studies, using molecular dynamics simulations, have started to shed light to the microscopic origins of internal friction. For example, Schulz *et al.* demonstrated that the internal friction, as well as the solvent friction, varies along the folding pathways.[69] Additionally, using extensive all-atom molecular dynamics simulations, Echeverria *et al.* were able to estimate the internal friction timescales for the cold shock protein under various solvent conditions[70] and found them to agree well with the corresponding experimental measurements.[15] Analysis of the reconfiguration dynamics of the unfolded chain further revealed that correlated hops in the dihedral space provide the dominant mechanism for internal friction of this protein.[70]

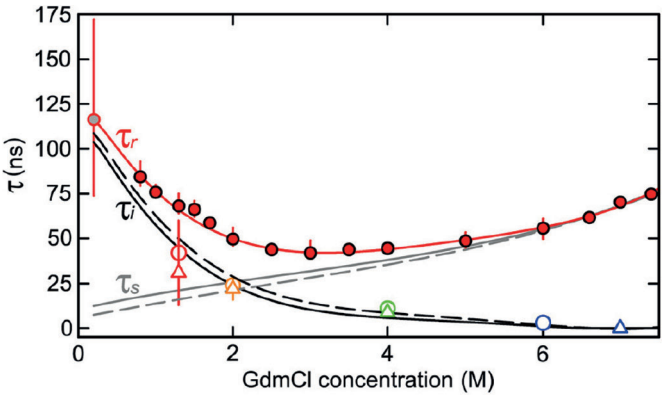


Figure 3. Determination of the contribution of internal friction to the reconfiguration time of unfolded proteins as a function of denaturant concentration (Guanidinium chloride (GdmCl)). Filled red circles correspond to the experimentally measured end-to-end reconfiguration times (τ_r). The solid (dashed) black line shows the timescale associated with internal friction (τ_i), calculated from the CRIF (Zimm) model. The solid (dashed) grey line shows the reconfiguration time expected for a CRIF (Zimm) model in the absence of internal friction (τ_s). The experimentally measured internal friction timescales are shown in open circles and triangles. These measurements show that, under low-denaturant concentration, internal friction plays a significant role in the dynamics of the unfolded state. Reproduced with permission from Ref. [15].

5. Connecting Simulations to Experiments and to Theory

Over the last decade, experimental methods, such as single molecule spectroscopy, have developed as important tools for characterizing the dynamics and structure of the unfolded ensemble (see reviews in Refs. [14,71–73]). Most importantly, these methods have allowed researchers to characterize the heterogeneity of the unfolded ensemble in a wide range of timescales,[14,34,74] ranging from nanoseconds to seconds. Additionally, recent advances in parallel algorithms, software, and hardware have made molecular dynamics simulations beyond a microsecond timescale accessible on many supercomputers and computer clusters. These developments have brought together a convergence of the timescales accessible through experiments and simulations, thus allowing the comparison between the independently obtained structural and kinetic information. This powerful approach of combining and complementing structural and dynamical information has been used to accurately determinate the ensembles of conformations sampled by unfolded proteins at the atomic level.[73,75]

Single molecule spectroscopy, based on fluorescence detection in combination with Förster resonance energy transfer (FRET), is a powerful tool for uncovering the conformational heterogeneity of the sampled sub-populations and for studying their interconversion dynamics. FRET methods, which are used as a spectroscopic ruler, can be applied to monitor conformational changes in the range of 10–80 Å, both at the steady state[72,76] and also in time-resolved fashion,[14,77] providing information about average intramolecular distances, the distance distributions, and the interconversion dynamics. A key advantage of these single molecule measurements is that they make it possible to extract dynamic information from equilibrium measurements.

Experimental information of the pairwise distances, obtained in FRET experiments, can be compared to distances obtained in molecular dynamics simulations or incorporated as constraints in simulations to bias the regions of sampled conformational space. In one approach, Nath *et al.*[78] used Monte Carlo simulations, constrained by

pairwise distance distributions from single-molecule FRET measurements, to characterize the conformation ensemble of IDPs α-Synuclein and Tau. Their work shows that few distance constraints (2 to 3) are sufficient to reproduce the global dimensions of the IDP's ensemble. Additionally, by modifying dihedral angle potentials, hydrogen bonding, effective residue-residue, and residue-environment interactions, the authors were able to tune the force field to reproduce the polymer scaling behavior and mean inter-residue distances obtained experimentally.

Inter-chain quenching experiments have also been used to characterize the unfolded ensemble.[79] Here, by analyzing the dynamics of contacts in the denatured Cytochrome (cb562), the authors determined that the contact formation rate followed a power-law dependence on the length of the probed segment of the polypeptide chain. Additionally, vibrational spectroscopy techniques have also been used to discriminate between different conformational manifolds sampled by unfolded and disordered peptides and proteins in solution.[80]

The translation of experimental measurements into structural information, as well as the connection between experimental and computational observables, is not necessarily straightforward and, in many cases, only qualitative comparisons are possible. In such cases, the concepts from statistical and polymer physics can provide the ideal bridge between single-molecule experiments and molecular dynamics simulations, setting a theoretical framework to describe the structural heterogeneity and dynamics of the unfolded ensemble. For example, the conformational dynamics of some unfolded polypeptides can be well described by random polymer models, such as the Rouse and Zimm models,[15,81,82] to predict, for example, the rate of contact formation or the reconfiguration times. Experimental studies of contact formation in unfolded proteins have found reasonable agreement with theoretical models, which can be further used to make predictions.[60, 65, 68, 79, 83]

As discussed above, single-molecule spectroscopy is ideally suited for probing the unfolded ensemble from a structural and dynamic viewpoints. This data can be complemented with the information available from other methods, such as NMR spectroscopy, where interatomic distances can be obtained from inter-proton distances derived from nuclear Overhauser effects (NOEs).[84–87] Additionally, in the presence of correlated motions, the joint distribution of interatomic distance can be complemented with the scalar or residual dipolar couplings obtained from NMR measurements to provide information about the orientations of the internuclear vectors, allowing to lift the degeneracy between ensembles with the same pairwise distance distributions, but with different structural conformations.[88] NMR methods have also been applied, under denaturing conditions, to define a set of representative conformations that can be used as the starting conformations for running unrestrained MD simulations.[33] In these studies, the combination of experimental and computational methods provide an unique framework to characterize the structural and physicochemical properties of the unfolded state of a protein in urea and to advance the understanding of the mechanisms of protein folding and unfolding.

One caveat of simulating unfolded proteins and IDPs is that the sampled ensembles are sensitive to the choice of protein force fields.[89] Most force fields have been parametrized to reproduce the folded state of proteins rather than the unfolded ensemble,[38,89] and therefore might fail to capture some important aspects of the latter's structural heterogeneity. Among the known discrepancies, currently used force fields tend to underestimate the average radius of gyration of the unfolded ensemble,[38] tend to be too hydrophobic[75] and tend to favor either α or β secondary structures. To address these pitfalls, one approach has been to optimize the commonly used force fields to reproduce the equilibrium conformational ensemble of short peptides by, for example, revising the backbone potentials[90,91] to reproduce the structural propensities of helix formation[90]. An alternative approach to characterize the conformational heterogeneity of unfolded proteins and IDPs is to use Monte Carlo simulations using the ABSINTH implicit solvation model,[92] instead of MD simulations. This force field has been designed primarily to determine the generic polymer character and conformational equilibria of IDPs in aqueous solutions by modeling the transfer of a polypeptide solute from the gas phase into a continuum solvent as the sum of a direct mean field interaction and a term to model the screening of polar interactions. The ABSINTH model has been used to study size distribution, long-range contacts, and secondary-structure formation of IDPs.[20,23]

6. Summary and Outlook

The characterization of the unfolded ensemble of proteins and IDPs at the atomic level resolution remains a challenging task, primarily because of the intrinsic conformational heterogeneity and complex dynamics of this phase. Hence, many difficulties remain in the ongoing quest to fully understand the structural and physicochemical properties of the unfolded state of proteins. For example, the reconfiguration dynamics of proteins with coil-like state conformations is mostly defined by the backbone dihedral transitions and the protein-solvent interactions. In contrast, in the collapsed state, besides backbone rearrangement transitions and protein solvent-interactions, the intramolecular dynamics may become dominated by the formation of secondary structure elements and native and non-native contacts. A molecular-level description of the extent of the contribution of each of these interactions,

and how they determine the roughness of the energy landscape, is still lacking.

The combination of experiments, simulations, and theory represents a promising approach that should enable us to gain further insights into the physicochemical interactions that determine the states populated by the unfolded ensemble, and the dynamic interconversion between these states. Furthermore, understanding the interactions that govern the unfolded ensemble should allow us to more finely distinguish between different degrees of disorder, or conformational heterogeneity. Here, particular emphasis should be placed on studying whether, and how, binding or post-translational chemical modifications regulate IDP function, by shifting the distribution of highly populated conformations. The related partial ordering phase transitions, where the final state still appears rather disordered (but less so than the initial state), need to be further probed, both computationally and experimentally.

Acknowledgements

This work was in part supported by the National Science Foundation (NSF) CAREER Award CHE-0846701 and by the University of Maryland.

References

[1] O. B. Ptitsyn, *Adv. Protein Chem.* **1995**, *47*, 83–229.
[2] D. Perl, C. Welker, T. Schindler, K. Schröder, M. A. Marahiel, R. Jaenicke, F. X. Schmid, *Nat. Struct. Mol. Biol.* **1998**, *5*, 229–235.
[3] A. Hoffmann, A. Kane, D. Nettels, D. E. Hertzog, P. Baumgärtel, J. Lengefeld, G. Reichardt, D. A. Horsley, R. Seckler, O. Bakajin, B. Schuler, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 105–110.
[4] A. Di Paolo, D. Balbeur, E. De Pauw, C. Redfield, A. Matagne, *Biochemistry* **2010**, *49*, 8646–8657.
[5] E. Sherman, G. Haran, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11539–11543.
[6] A. Yu Grosberg, A. R. Khokhlov, *Statistical Physics of Macromolecules*, American Institute of Physics, New York, **1994**.
[7] C. J. Camacho, D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6369–6372.
[8] P. G. Wolynes, J. N. Onuchic, D. Thirumalai, *Science* **1995**, *267*, 1619–1620.
[9] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, N. D. Socci, *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3626–3630.
[10] M. Sadqi, L. J. Lapidus, V. Muñoz, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12117–12122.
[11] W. Y. Yang, M. Gruebele, *Nature* **2003**, *423*, 193–197.
[12] J. Kubelka, J. Hofrichter, W. A. Eaton, *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88.
[13] D. Nettels, I. V Gopich, A. Hoffmann, B. Schuler, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2655–2660.
[14] B. Schuler, H. Hofmann, *Curr. Opin. Struct. Biol.* **2013**, *23*, 36–47.
[15] A. Soranno, B. Buchli, D. Nettels, R. R. Cheng, S. Müller-Späth, S. H. Pfeil, A. Hoffmann, E. A. Lipman, D. E. Makarov, B. Schuler, *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17800–17806.
[16] A. Borgia, B. G. Wensley, A. Soranno, D. Nettels, M. B. Borgia, A. Hoffmann, S. H. Pfeil, E. A. Lipman, J. Clarke, B. Schuler, *Nat. Commun.* **2012**, *3*, 1195.
[17] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, A. K. Dunker, *Proteins: Struct. Funct. Bioinf.* **2001**, *42*, 38–48.
[18] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, *J. Mol. Biol.* **2004**, *337*, 635–645.
[19] K. A. Dunker, I. Silman, V. N. Uversky, J. L. Sussman, *Curr. Opin. Struct. Biol.* **2008**, *18*, 756–764.
[20] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, R. V. Pappu, *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183–8188.
[21] R. B. Best, J. Mittal, *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 11087–11092.
[22] N. Lyle, R. K. Das, R. V. Pappu, *J. Chem. Phys.* **2013**, *139*, 121907.
[23] W. Meng, N. Lyle, B. Luan, D. P. Raleigh, R. V. Pappu, *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 2123–2128.
[24] G. Parisi, *Phys. Rev. Lett.* **1983**, *50*, 1946–1948.
[25] R. A. Goldstein, Z. A. Luthey-Schulten, P. G. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918–4922.
[26] J. D. Bryngelson, J. N. Onuchic, N. Socci, P. G Wolynes, *Proteins: Struct. Funct. Bioinf.* **1995**, *21*, 167–195.
[27] J. N. Onuchic, P. G. Wolynes, *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.
[28] D. A. Potoyan, G. A. Papoian, *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17857–17862.
[29] N.-j. Deng, W. Dai, R. M. Levy, *J. Phys. Chem. B* **2013**, *117*, 12787–12799.
[30] D. Shortle, M. S. Ackerman, *Science* **2001**, *293*, 487–489.
[31] B. Zagrovic, C. D. Snow, S. Khaliq, M. R. Shirts, V. S. Pande, *J. Mol. Biol.* **2002**, *2836*, 153–164.
[32] C. M. Dobson, *Nature* **2003**, *426*, 884–890.
[33] M. Candotti, S. Esteban-Martín, X. Salvatella, M. Orozco, *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 5933–5938.
[34] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, B. Schuler, *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155–16160.
[35] R. K. Das, R. V. Pappu, *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392–13397.
[36] G. A. Papoian, P. G. Wolynes, *Biopolymers* **2003**, *68*, 333–349.
[37] B. E. Bowler, *Curr. Opin. Struct. Biol.* **2012**, *22*, 4–13.
[38] K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, D. E. Shaw, *J. Am. Chem. Soc.* **2012**, *134*, 3787–3791.
[39] S. Tripathi, G. I. Makhatadze, A. E. Garcia, *J. Phys. Chem. B* **2012**, *117*, 800–810.
[40] V. A. Voelz, M. Jager, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, S. Weiss, V. S. Pande, *J. Am. Chem. Soc.* **2012**, *134*, 12565–12577.
[41] V. Ozenne, J. K. Noel, P. O. Heidarsson, S. Brander, F. M. Poulsen, M. R. B. Jensen, B. B. Kragelund, M. Blackledge, J. Danielsson, *J. Mol. Biol.* **2013**, 1–13.
[42] G. G. Maisuradze, A. Liwo, S. Odziej, H. A. Scheraga, *J. Am. Chem. Soc.* **2010**, *132*, 9444–9452.
[43] H. J. Dyson, P. E. Wright, *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
[44] P. E. Wright, H. J. Dyson, *Curr. Opin. Struct. Biol.* **2010**, *19*, 31–38.

[45] G. A. Papoian, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14237–14238.
[46] P. I. Zhuravlev, G. A. Papoian, *Q. Rev. Biophys.* **2010**, *43*, 295–332.
[47] P. Tompa, M. Fuxreiter, *Trends Biochem. Sci.* **2008**, *33*, 2–8.
[48] D. Vuzman, Y. Levy, *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 21004–21009.
[49] P. Tompa, *Curr. Opin. Struct. Biol.* **2011**, *21*, 419–425.
[50] J. Chen, *Arch. Biochem. Biophys.* **2012**, *54*, 123–131.
[51] Y. Levy, P. G. Wolynes, J. N. Onuchic, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 511–516.
[52] J. Mittal, T. H. Yoo, G. Georgiou, T. M. Truskett, *J. Phys. Chem. B* **2013**, *117*, 118–124.
[53] D. A. Potoyan, G. A. Papoian, *J. Am. Chem. Soc.* **2011**, *133*, 7405–7415.
[54] M. S. Cheung, D. Klimov, D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4753–4758.
[55] J. Mittal, R. B. Best, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20233–20238.
[56] E. Rivera, J. Straub, D. Thirumalai, *Biophys. J.* **2009**, *96*, 4552–4560.
[57] M. E. McCully, D. A. C. Beck, V. Daggett, *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17851–17856.
[58] V. Muñoz, *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 395–412.
[59] R. Zwanzig, *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2029–2030.
[60] R. Best, G. Hummer, *Phys. Rev. Lett.* **2006**, *96*, 228104.
[61] R. B. Best, G. Hummer, *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 1088–1093.
[62] L. L. Chavez, J. N. Onuchic, C. Clementi, *J. Am. Chem. Soc.* **2004**, *126*, 8426–8432.
[63] D. E. Sagnella, J. E. Straub, D. Thirumalai, *J. Chem. Phys.* **2000**, *113*, 7702–7711.
[64] J. J. Portman, S. Takada, P. G. Wolynes, *J. Chem. Phys.* **2001**, *114*, 5082–5096.
[65] M. Buscaglia, L. J. Lapidus, W. A. Eaton, J. Hofrichter, *Biophys. J.* **2006**, *91*, 276–288.
[66] T. Cellmer, E. R. Henry, J. Hofrichter, W. A. Eaton, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18320–18325.
[67] A. Sekhar, P. Vallurupalli, L. E. Kay, *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 19268–19273.
[68] R. R. Cheng, A. T. Hawk, D. E. Makarov, *J. Chem. Phys.* **2013**, *138*, 074112.
[69] J. C. F. Schulz, L. Schmidt, R. B. Best, J. Dzubiella, R. R. Netz, *J. Am. Chem. Soc.* **2012**, *134*, 6273–6279.
[70] I. Echeverria, D. E. Makarov, G. A. Papoian, **2014**, *J. Am. Chem. Soc.* **2014**, *136*, 8708–8713.
[71] M. Vendruscolo, *Curr. Opin. Struct. Biol.* **2007**, *17*, 15–20.
[72] B. Schuler, W. A. Eaton, *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26.
[73] L. J. Lapidus, *Curr. Opin. Struct. Biol.* **2013**, *23*, 30–5.
[74] H. Oikawa, Y. Suzuki, M. Saito, K. Kamagata, M. Arai, S. Takahashi, *Sci. Rep.* **2013**, *3*, 2151.
[75] K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, W. A. Eaton, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1528–1533.
[76] B. Schuler, E. A. Lipman, W. A. Eaton, *Nature* **2002**, *419*, 743–748.
[77] L. Albizu, M. Cottet, M. Kralikova, S. Stoev, R. Seyer, I. Brabet, T. Roux, H. Bazin, E. Bourrier, L. Lamarque, C. Breton, M.-L. Rives, A. Newman, E. Trinquet, M. Manning, J.-P. Pin, *Nat. Chem. Biol.* **2012**, *6*, 587–594.
[78] A. Nath, M. Sammalkorpi, D. C. DeWitt, A. J. Trexler, S. Elbaum-Garfinkle, C. S. O'Hern, E. Rhoades, *Biophys. J.* **2012**, *103*, 1940–1949.
[79] N. D. B. Ford, D.-W. Shin, H. B. Gray, J. R. Winkler, *J. Phys. Chem. B* **2013**, *117*, 13206–13211.
[80] R. Schweitzer-Stenner, *J. Phys. Chem. B* **2013**, *117*, 6927–6936.
[81] D. E. Makarov, K. W. Plaxco, *J. Chem. Phys.* **2009**, *131*, 085105.
[82] D. E. Makarov, *J. Chem. Phys.* **2013**, *138*, 014102.
[83] L. Milanesi, J. P. Waltho, C. A. Hunter, D. J. Shaw, G. S. Beddard, G. D. Reid, S. Dev, M. Volk, *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 19563–19568.
[84] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, C. M. Dobson, *J. Am. Chem. Soc.* **2005**, *127*, 476–477.
[85] T. L. Religa, J. S. Markson, U. Mayor, S. M. V. Freund, A. R. Fersht, *Nature* **2005**, *437*, 1053–1056.
[86] H. Heise, S. Luca, B. L. de Groot, H. Grubmüller, M. Baldus, *Biophys. J.* **2005**, *89*, 2113–2120.
[87] T. Mittag, J. D. Forman-Kay, *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
[88] G. Bouvignies, P. Bernadó, S. Meier, K. Cho, S. Grzesiek, R. Brüschweiler, M. Blackledge, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885–13890.
[89] S. Piana, K. Lindorff-Larsen, D. E. Shaw, *Biophys. J.* **2011**, *100*, L47–L49.
[90] R. B. Best, J. Mittal, *J. Phys. Chem. B* **2010**, *114*, 8790–8798.
[91] P. S. Nerenberg, T. Head-Gordon, *J. Chem. Theory Comput.* **2011**, *7*, 1220–1230.
[92] A. Vitalis, R. V. Pappu, *J. Comput. Chem.* **2008**, *30*, 673–699.

Received: January 27, 2014
Accepted: March 19, 2014
Published online: August 18, 2014