Molecular Simulations of Protein Disorder

A Review

Sarah Rauscher [1,2] and Régis Pomès[1,2,*]

1. Molecular Structure and Function, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G1X8

2. Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, ON, Canada, M5S 1A8

* To whom correspondence should be addressed. Email: pomes@sickkids.ca.

Date: ***, 2009

Submitted to *Biochemistry and Cell Biology*

Frequently Used Abbreviations:

IDP: intrinsically disordered protein, IDR: intrinsically disordered region, IUP: intrinsically unstructured peptide, MD: molecular dynamics, STDR: simulated tempering distributed replica sampling, PRE: paramagnetic relaxation enhancements, NMR: nuclear magnetic resonance

Abstract:

Protein disorder is abundant in genomes throughout all kingdoms of life, and serves many biologically important roles. Disordered states of proteins are challenging to study experimentally due to their structural heterogeneity and tendency to aggregate. Molecular simulation is ideally suited to characterize the conformational ensembles of intrinsically disordered proteins. The application of simulation to the study of disordered states is relatively recent, and is described in this review in the context of experimental and bioinformatics approaches. Different levels of detail are possible in simulations, from coarse-grained lattice models to all-atom representations with explicit water. We address the challenges inherent to simulating disordered states with examples of studies from the literature and our own work. Importantly, we highlight the interdisciplinary nature of the study of disordered proteins. Experimental data can be incorporated into simulations, and simulations can provide predictions for experiment. In this way, simulations have been integrated seamlessly into the already-existing methodologies for the study of disordered state ensembles. This review is intended as both an overview and a guide for structural biologists seeking accurate, atomic-level descriptions of disordered state ensembles.

INTRODUCTION

Proteins can simplistically be classified into two categories: those that fold (*folders*) and those that don't (*non-folders*). Historically, the structure-function paradigm has defined our understanding of proteins. It was thought that a protein's function is 'encoded' in the structure of its folded, and therefore biologically active, state.[1] This view was supported by numerous experiments and theories, including Fischer's lock and key model.[2] Anfinsen's classic experiment demonstrated that the information defining the secondary and tertiary structure of proteins resides in the amino acid sequence.[3] A protein's loss of function in the denatured state was associated with a loss of structure.[4] Structural biologists therefore focussed their studies on proteins with well-defined folded states, and proteins capable of functioning without a unique folded state were thought to be rare.[1, 5] The connection between sequence, structure and function is essential to the understanding of enzymes and transport proteins. In contrast, the functions of many proteins involved in cell signalling, molecular recognition and transcriptional regulation rely on the presence of unstructured regions.[6-8] Intrinsically disordered peptides (IDPs) lack a unique, folded conformation under physiological conditions.[1] In the past decade, the structure-function paradigm has been revisited in light of the prevalence of IDPs.[9, 10]

Disordered proteins are both abundant in nature and diverse in their functions,[11] underscoring the need for detailed studies. Bioinformatics approaches predict that disordered sequences are prevalent in all genomes, especially in eukaryotes, for which long disordered regions (> 30 residues) are predicted to be present in 33 % of all proteins.[7] In mammals, 75 % of signalling proteins are predicted to contain long disordered regions.[6] Disorder has been demonstrated to correlate with disease-related proteins (*disorder in disorders* – the D^2 concept).[12] Importantly, 79 % of cancer-associated proteins are predicted to contain contiguous regions of disorder longer than 30 residues. Disordered proteins therefore are an emerging class of drug targets.[13] The biological importance of disorder has driven the development of computational and experimental approaches to characterize IDPs, which can be significantly more challenging than studying proteins with well-defined folded states.[14] Biophysical studies of disordered states are complicated by structural heterogeneity, and consequently the need to describe a dynamic ensemble rather than a singular structure.[15]

Defining Disorder

Before we begin, we describe the nomenclature of the study of disordered proteins including definitions of the terms *intrinsically disordered proteins (IDPs), intrinsically disordered regions (IDRs)*, *configuration*, *conformation*, *molten globule, premolten globule, random coil* and *folded state*. Intrinsically disordered proteins

(IDPs) are proteins that lack a well-defined, rigid structure along their entire length. Segments that are unstructured both in isolation and in the context of the full length protein to which they belong are referred to as intrinsically disordered regions (IDRs).[5] IDPs are globally disordered, while IDRs exhibit local disorder.[1] Proteins populating many dissimilar conformations may be thought of as the "liquid state" of proteins.[16]

An important distinction is that between configurations and conformations.[17] The *configuration* of a peptide is defined by the three-dimensional coordinates of each of its atoms. Configuration space is, in principle, infinite and continuous. A *conformation* is the set of configurations which are related to each other based on a selected measure of structural similarity. Conformational space is therefore discrete, and is defined by the criterion used to distinguish between conformations. Clustering is often employed for the purpose of defining conformations.[17] A thermodynamically-accessible conformation is a conformation which is populated under a particular set of conditions, such as temperature and concentration.[17] In order to fully describe a disordered state ensemble, in principle, one should specify each conformation along with its population.[14]

It has been suggested that globular proteins can access at least four unique states: native, molten globule, pre-molten globule and unfolded.[9] The *folded state* corresponds to a conformation that is both highly populated and structured,[18] which may also be referred to as *native state* if it is the biologically relevant and experimentally observable conformation.[17] All folded proteins exist in equilibrium with their unfolded states.[19] Relative to the folded state, the *molten globule* exhibits an increase in the hydrodynamic volume of no more than 50 %.[9] Its secondary structure and folding pattern are native-like,[9] with loosened or 'molten' tertiary interactions.[1, 20] In the *premolten globule* state, there is no rigid tertiary structure, approximately 50 % native-like secondary structure and a hydrodynamic volume three times larger than the folded state.[9] The central tenet of the Protein Trinity, consisting of the folded state, the molten globule and the random coil[21] and the Protein Quartet, which includes the pre-molten globule as a unique thermodynamic state,[9] is that any of these states may be referred to as the *native state*, which is the state relevant to a protein's biological function. Transitions between any of these states may be functionally important.[9, 21] There is also an important distinction to be made between the unfolded state and the denatured state, which is a chemically-stabilized or temperature-induced unfolded state.[22] In the *random coil* state, the backbone conformation of each residue is independent of the φ and ψ angles of neighbouring residues.[23] Polypeptides cannot be described as true random coils since each φ,ψ pair is not independent, resulting in preferred and sterically-disfavoured conformations.[23, 24]

Unstructured protein ensembles and the relationships between them are not completely understood,[6] and there are different degrees of conformational disorder. It has been proposed that IDPs should in fact be

subdivided into two classes: intrinsic coils and premolten globules.[9] 'Intrinsically disordered' is used to describe any state that is incompletely folded, and 'intrinsically unstructured' to describe random-coil-like or pre-molten globule-like states.[13] While it is useful to conceptualize in terms of these discrete states, it is important to remember that there is in fact a continuum of disordered states, with varying degrees of compactness and varying amounts of secondary and tertiary structure.[25-27]

Of particular interest are proteins that are natively unfolded, for which the native state corresponds to an unstructured ensemble. We introduce two terms: *folder* and *non-folder*. It is necessary to distinguish proteins that are capable of folding under certain conditions (folders) from those which do not have a folded state, or ordered state, under any known conditions (non-folders). Most experimental techniques are suited for characterizing ordered structure, that is, to characterize the folded state. The 'native state' of many proteins *in vivo* may not be folded, but instead consist of an ensemble of interconverting structures, none of which correspond to a folded state. In this review, we classify IDPs that are competent to fold upon binding a partner as 'folders'. Only IDPs that possess no folded state under any known conditions are referred to as 'non-folders'. Many IDPs/IDRs fold upon binding to their targets.[28] Molecular recognition is analogous to protein folding, since both processes have a folded state which is thermodynamically stable relative to the unfolded state of higher conformational entropy.[29]

Biological Functions of Disorder

Disordered proteins are essential in a wide variety of cellular activities, which can be separated into two main functional classes: recognition and entropic chains.[8, 30] To motivate the importance of studying IDPs/IDRs, we provide a few select examples, noting that there are several excellent reviews on the topic.[1, 6, 8, 13, 26, 28, 30] The functions of disordered proteins are complementary to those of proteins with well-defined three dimensional structures. One functional niche of IDPs/IDRs is induced folding upon binding to a partner, which may be another protein, a nucleic acid, a membrane or a small molecule.[12] Such disorder-to-order transitions are important to both the DNA-binding domains and transcriptional activation domains of eukaryotic transcription factors.[1] There are several examples of DNA binding domains with IDRs that undergo local folding transitions upon binding to DNA, resulting in sequence specific recognition. It is also common for transcriptional activation domains to be unstructured, with 'induced folding' only in the presence of their targets. Similarly, proteins involved in RNA recognition become structured in the presence of their cognate RNA partner.[1]

There are several biological advantages of binding involving a disordered partner versus the binding of two structured partners. The absence of a specific, ordered structure facilitates the possibility of binding to multiple targets.[1] Conformational flexibility may confer a functionally important binding 'promiscuity', with different conformations suited to different binding targets. This is referred to as 'one-to-many' signalling.[13] Conversely, many different disordered regions may bind to a common partner in 'many-to-one' signalling.[13] Proteins with multiple binding partners are hubs in protein interaction networks. Disordered proteins frequently function as hubs or interaction partners of hubs due to their intrinsic ability to participate in one-to-many and many-to-one signalling.[13]

An important advantage of binding a disordered partner is high specificity with modest affinity. The loss in conformational entropy due to the disorder-to-order transition must be compensated for by favourable interactions at the binding interface. The result is an interaction with high complementarity and low binding affinity, allowing the complex to readily dissociate and terminate the signalling event. Binding can be readily modulated by cellular signals, including post-translational modifications.[1] Protein-protein interactions are often important drug targets, and disordered protein interactions may be easier to block with competitors compared to a similar interaction of two structured proteins due to their reduced affinity.[13]

There may also be functionally important disorder in the bound state, in so-called "fuzzy complexes".[13, 31] Sic1 is an example of an IDP that is largely unstructured while performing its biological function.[25] It is an inhibitor of cyclin-dependent kinases and is involved in regulation of the yeast cell cycle. Its flexibility enables transient interactions in a dynamic complex, with each binding site interacting in a dynamic equilibrium. Sic1 does not exhibit a disorder-to-order transition.[32, 33] Its structure resembles a molten globule or pre-molten globule, with some secondary and tertiary structure.[34] Native molten globules represent a disordered yet functional state.[35]

Entropic chains represent a subset of IDPs that are unstructured under physiological conditions and do not undergo a disorder-to-order transition.[30] Their functions are therefore completely outside the realm of proteins with well-defined structures.[36] Flexible linkers or spacers connect domains and allow them to move with respect to each other.[36] Entropic bristles exclude volume due to thermal motion of the polypeptide. They prevent the motion of larger molecules, while allowing water and small solutes to pass through.[37] Elastomeric proteins, including elastin, spider silk and resilin, are entropic chains with a function that necessitates disorder.[38] Their high configurational entropy results in elastic recoil in two possible biological roles: 'molecular springs' or 'shock absorbers'.[39]

Unstructured regions are often found in proteins that are targeted for degradation, and their turnover facilitates rapid cellular response to changing conditions.[1] Protease degradation is used to regulate concentrations of certain proteins in the cell.[1] The cell cycle must be sensitive to environmental conditions. It is thought that the rapid turnover of IDPs by the cell's proteolytic machinery may play a role in this sensitivity, and the necessary short response times. The abundance and residence times of IDPs are tightly regulated in the cell and their overexpression or underexpression are associated with several diseases including cancer, Alzheimer's disease and Parkinson's disease.[40] In general, the half-lives of disordered proteins are shorter than those of folded proteins, but can be extended by modifications like phosphorylation or binding to partners.[11]

There is a major biological disadvantage to disordered proteins. IDPs/IDRs typically have exposed hydrophobic side chains in their partially folded or completely unstructured conformational ensembles, which may predispose these proteins to forming aggregates.[41] IDPs are implicated in the pathogenesis of several protein misfolding diseases, including Alzheimer's disease (amyloid β-protein), Parkinson's disease (α-synuclein), type II diabetes (amylin) and Huntington's disease (polyglutamine repeats).[42] Aggregation of IDPs is also a matter of practical interest due to the challenges of applying biophysical approaches to insoluble aggregates *in vitro*.[41] In general, the formation of amyloid fibrils requires a partially or fully unfolded precursor.[42, 43] The compact native state of folded proteins must be at least partially unfolded to form states that are competent for aggregation, with the main chain and hydrophobic groups exposed.[43] IDPs do not need to undergo such conformational rearrangements to form aggregation-prone states. They are therefore at risk for self-aggregation, and represent a significant portion of known amyloidogenic proteins.[42] Elastomers represent a class of disordered proteins that must aggregate to function. Their high combined content of proline and glycine preclude the formation of amyloidogenic aggregates and result in the disorder necessary for elastic recoil.[38]

Sequence Features of Disordered Proteins – Bioinformatics Approaches

The impressive biological importance of disorder has prompted several groups to develop techniques to predict disorder based exclusively on amino acid sequence. There are relatively few well-characterized examples of disordered proteins, and bioinformatics has therefore been useful in demonstrating that these few examples are part of a large set of IDPs found in genomes throughout all kingdoms of life.[13] Just as the primary sequence encodes the folded state of proteins, it similarly defines the free energy landscape, including the unfolded state. For proteins with no folded state or ordered conformation, the primary sequence precludes the possibility of folding. It is therefore expected that the amino acid sequence should encode a protein's folded state, as well as

a protein's inability to fold. The fact that predictors have been developed that can discriminate between ordered and disordered proteins based on sequence alone supports this idea.[6, 44]

A significant limitation to the prediction of disorder is the presence of ordered sequences misclassified in databases of disordered proteins.[45] When databases of ordered and disordered proteins are used as training sets for disorder predictors, false positives in both data sets limit the accuracy of the resulting predictors.[6] 'Noise' in the disordered protein data set is also partly due to inconsistency in the definition of structural disorder. Different methods have different limitations with regard to identifying structural disorder.[44] For example, CD may misclassify a structured loop as being disordered. Crystal packing in x-ray crystallography results in the ordering of segments that are disordered in solution. Poor signal dispersion in NMR spectroscopy is often used to classify a protein as disordered, which is not sufficient to distinguish between molten globule and coil states.[44]

Despite these limitations, several groups have identified sequence signatures of disorder. Intrinsically disordered proteins and globular proteins have been distinguished by correlating mean hydrophobicity with net charge.[46] A set of natively-unfolded proteins have lower mean hydrophobicity and higher net charge compared to a set of small, monomeric globular proteins. This separation is due to a combination of two effects. First, increased charge repulsion compared to globular proteins favours unfolding. Second, there are fewer hydrophobic groups, and therefore less driving force for hydrophobic collapse.[46] Since this observation was first made, there have been many more disordered sequences confirmed experimentally. We have therefore analysed the hydrophobicity and net charge of the disordered regions in the database of disordered proteins (DISPROT)[47] in its present state (version 4.9). In Figure 1a, hydrophobicity versus net charge is plotted for disordered regions from DISPROT, and a set of ordered proteins from the Protein Data Bank (PDB). The average hydrophobicity is lower and the average net charge is higher for disordered sequences compared to ordered sequences. However, it is not possible to unambiguously classify a sequence as disordered based on these properties alone. This analysis is consistent with a recent study showing that ordered and disordered sequences cannot be unambiguously distinguished based on net charge and mean hydrophobicity. In fact, the charge-hydrophobicity plot is only a two-dimensional projection of the twenty-dimensional amino acid composition space. [48]

Intrinsically disordered peptides also have unique amino acid compositions. There is a clear enrichment and depletion of certain amino acids relative to the compositions of ordered proteins.[49] We have computed the average composition of disordered proteins in the current version of the DISPROT database relative to the set of

structured proteins from the PDB. The amino acids that are enriched in disordered sequences are: P, Q, S, E, G, K, D, R and A. There is a decrease in the composition of T, N, M, H, V, F, L, Y, W, C and I. Charged residues are enriched, while large hydrophobic and aromatic residues are depleted in disordered sequences. This is consistent with a previous analysis of disordered region amino acid compositions in which it was found that W, C, F, I,Y, V, L, and N are depleted while A, R, G, Q, S, P, E, and K are enriched.[49] A support vector machine trained on disordered and ordered sequences incorporating amino acid composition alone can recognize disordered sequences with an accuracy of 87 %.[50] Such a high accuracy suggests that composition is a very important determinant of disorder. Interestingly, a high combined composition of proline and glycine is a necessary requirement for elastomeric proteins, which represent an important class of disordered proteins.[38] Conservation of sequence composition may be more important for disordered regions than conservation of exact sequence motifs.[6] Sequence composition alone, however, is insufficient to conclusively determine if a sequence is disordered.[48]

Disordered sequences are significantly less complex than the sequences of globular proteins. The complexity, as measured by Shannon's entropy, of natively-folded proteins is similar to that of random sequences. The complexity of non-globular, disordered proteins is much lower. Disordered proteins, which do not need to maintain a particular folded structure to function, have fewer sequence constraints and low complexity may be a consequence. It is important to note that there is no observed upper limit on the complexity of IDPs, since a high sequence complexity does not guarantee that a protein will fold.[45] Low sequence complexity is also not a sufficient condition for disorder. There are low complexity sequences that have repetitive ordered structures, such as fibrous proteins, collagen and coiled coils.[45, 51] Reduced complexity sequences do not exclusively lead to disorder. In fact, nearly 20 % of proteins in the human genome contain multiple repeats of 30-40 residues. Many of these are ankyrin repeat proteins, which are stable and cooperatively folded.[52] The reduced sequence complexity of IDPs has particularly important implications for the simulation of these sequences. Statistical averaging in MD simulations is enhanced by repetitive sequences, since each repeat motif is simulated several times in the same simulation.[38]

The IUPred disorder predictor computes a pairwise interaction energy based on amino acid composition as a test of a sequence's 'foldability' and does not rely on databases of disordered proteins.[44, 53] If the pairwise energy is below a certain threshold, the sequence is capable of forming an ordered structure. Using this predictor, IDPs/IDRs are distinguished from globular proteins on an energetic basis. The success of IUPred as a disorder predictor suggests that the compositions of disordered proteins are biased to form fewer favourable interactions compared to globular proteins. It is important to note that random sequences with low predicted

pairwise energy content do not necessarily fold. IUPred predicts coil-like disordered sequences to be energetically 'neutral', with nearly balanced attractive and repulsive interactions, while molten globule-like IDPs have a net energetic stabilization.[44]

There are currently at least 25 disorder predictors, and the accuracy of disorder prediction continues to increase.[13, 54] Currently, no disorder predictor is completely accurate on its own, but their accuracy can be improved by combining multiple approaches.[27, 55] Disordered regions can be predicted using a metaserver simultaneously utilizing multiple disorder predictors (DISMETA).[56] Disordered sequences are predicted to be more prevalent in the genomes of eukaryotes compared to prokaryotes.[6] More than 20 % of eukaryotic proteins are predicted to have a majority of residues in disordered regions.[57] Around 25 % of all proteins are predicted to have some disorder.[10] Given the huge number of IDPs and IDRs predicted by bioinformatics, an ongoing challenge is to confirm and characterize these disordered sequences using both experiment and simulation.

Identifying and Characterizing Disordered Proteins -- Experimental Approaches

In the timeline of biochemistry, the discovery of IDPs and IDRs is relatively recent. Classical biochemical techniques are biased towards the identification and characterization of folded proteins. A function can now be mapped to a particular gene and the corresponding protein is produced, purified and studied using a multitude of approaches.[26] NMR spectroscopy has so far proven to be the most effective experimental approach for obtaining high resolution, site-specific structural information for both IDPs and IDRs.[5, 26] Heteronuclear multidimensional NMR is useful for studying disordered states.[9] Residual secondary structure in disordered regions is usually transient, and confined to short segments.[5] Secondary chemical shifts provide site-specific information, which can be used to quantify fractional secondary structure propensity.[58] Paramagnetic relaxation enhancement (PRE) is useful for observing long range contacts.[5] Excellent reviews of the application of NMR to disordered and unfolded states are available.[5, 10, 20, 25]

Other techniques used to identify disordered regions include circular dichroism, hydrodynamic measurements, susceptibility to proteolytic degradation, electrospray ionization mass-spectrometry, proton-deuterium exchange methods and fluorescence spectroscopy.[26, 54, 59] Missing electron density in a structure determined by x-ray crystallography is often interpreted as evidence of a disordered region. Flexibility leads to incoherent x-ray scattering, resulting in missing electron density in a crystal structure.[9] It is still not completely clear whether proteins that are characterized as disordered *in vitro* remain unstructured *in vivo*,[11] and some IDPs may become more structured in the crowded environment of the cell.[28] Binding-induced folding of IDPs can be

observed using fluorescence, with implications for the design of protein biosensors.[60] Förster resonance energy transfer (FRET) can also be used to obtain a distribution of distances between two residues with fluorophores (accessible distances in the range of 10 to 80 Å).[14, 25] These distance distributions can be incorporated as restraints in the determination of conformational ensembles,[5] or compared to distances from simulations.[61] Several other experimental approaches can provide quantitative information for simulation restraints in order to represent a disordered ensemble of conformations consistent with the data.[25]

Hybrid Approaches: Using Experimental Restraints in Simulations of Conformational Ensembles

Experimental measurements of disordered states often represent averages over a broad ensemble of unrelated structures. Ensemble averages do not provide information about the underlying conformational distribution. Several approaches have been developed to impose experimentally-derived restraints on ensembles of conformations. One of the most straightforward approaches involves the generation of a set of random structures which are population-weighted to create an ensemble consistent with a set of restraints.[62-64] The ENSEMBLE algorithm makes use of a wide variety of experimental restraints such as chemical shifts, nuclear Overhauser effects (NOEs), PREs, residual dipolar couplings, hydrogen exchange protection factors, solvent-accessible surface area and hydrodynamic radius.[63, 64] The goal of ENSEMBLE is to generate the simplest ensemble of conformations that is consistent with the input data in order to avoid overfitting.[64] A conformer pool of sterically-plausible unfolded conformations can be generated using TraDES.[65] In principle, any method of unbiased conformational sampling can be used to generate a conformer pool, including molecular dynamics (MD) simulations.[63] An iterative process is followed by applying subsets of the experimental restraints to the conformations selected by ENSEMBLE,[25, 66] or through iterative conformational sampling.[64]

An analogous approach is the ensemble optimization method (EOM).[67] A conformer pool is generated with the flexible-meccano algorithm, which randomly selects dihedral angles for each residue from a library of coil conformations.[68] An x-ray scattering curve is calculated for each conformer and sets of conformers that fit the experimental SAXS measurements are selected using a genetic algorithm.[67] Different EOM runs produce different conformer pools, all of which match the same experimental SAXS profile.[67] This is because SAXS is a low-resolution method.[67] Completely different conformational ensembles can be consistent with a given set of experimental data.[69]

Even with the extensive set of experimental restraints currently available, there is still insufficient information to uniquely determine disordered ensembles.[64] It is an underdetermined problem because the

number of degrees of freedom in the system far exceeds the number of experimental restraints.[69] Sparse experimental data can be complemented by the information contained in molecular mechanics force fields.[14] Experimentally-derived restraints can be combined directly with all-atom MD simulations to produce disordered state ensembles that are consistent with experimental data.[14] There have been many such hybrid experimental-MD simulation studies using different sets of experimentally-derived restraints.

Chemical shift information can be directly incorporated as restraints, or used as a criterion to evaluate the accuracy of a conformational ensemble. For instance, the partially-unfolded signalling state of photoactive yellow protein was modeled using MD simulations with restraints based on NMR chemical shifts. This approach is effective when the partially unfolded state represents a relatively small perturbation from the folded state, and the structure of the folded state is known.[70] Another approach that makes use of chemical shift information is the energy-minima mapping and weighting (EMW) method.[71, 72] The EMW method consists of three steps: conformational sampling, ensemble generation and validation. This method is motivated by the idea that a disordered ensemble is a set of energetically favourable conformations. To ensure that a structurally heterogeneous set of initial conformations is generated, EMW uses a combination of high temperature MD and end-to-end distance restraints.[71] Conformations are then simulated at low temperature and energy-minimized. Using the library of initial conformations, EMW generates many candidate ensembles, each of which is independently consistent with a set of $^{13}C\alpha$ chemical shifts. Validation of the ensembles is performed using ^{13}CO, ^{15}N, and $H\alpha$ chemical shifts and scalar J-couplings.[72] The EMW method has been used to study the conformational ensemble of a C-terminal fragment of p21. p21 is an IDP capable of recognising and binding to at least 25 different substrates. Its structural plasticity allows it to adopt both α-helical and extended structures. Interestingly, the bound conformations of p21 were found to exist in the conformational ensemble of its unbound state, supporting a conformational selection mechanism for p21's binding promiscuity.[72]

MD simulations can use time-averaged restraints or apply restraints simultaneously to an ensemble of simulations run in parallel.[73-75] Time-averaged restraints minimize the disturbance to the force field due to the addition of an artificial term. The ensemble of structures from the complete simulation trajectory satisfies all restraints, while no single structure is required to do so.[73] For example, MD simulations of the XAO peptide (Ac-$XX(A)_7OO$-NH_2, X is diaminobutyric acid and O is ornithine) using simulated annealing were performed.[76] Inter-proton distances from rotating frame nuclear Overhauser effect (ROE) intensities and dihedral angles from scalar couplings were included as time-averaged restraints. The structure of the XAO peptide was found to be an interconverting ensemble, with no extend polyproline II (PPII) structure. Although individual residues of the XAO peptide sample the PPII conformation, the PPII helix is not the unique conformation of the XAO peptide.[76]

Importantly, the properties of the ensemble were verified by agreement with the radius of gyration from SAXS measurements.[76] Using the method of local elevation sampling, ^{3}J coupling constants can be adaptively restrained by keeping track of conformations sampled throughout the simulation.[77]

The Monte Carlo Replica Sampling (MCRES) method[75] enforces experimentally-derived distance restraints on an ensemble of replicas simulated in parallel. Restraints are imposed using an energetic penalty on the ensemble-average of distances, not on any individual structure. PRE measurements provide ensemble-averaged measurements of distances using spin labels.[75] The distance range of the PRE method (~20 Å) exceeds that of NOE distance measurements (~ 5 Å) and is therefore well-suited to detect the long range, transient contacts in disordered ensembles.[15] The denatured state of bovine acyl-coenzyme A binding protein (ACBP) was studied using MCRES with a Cα-representation of the protein.[75] For this system, twenty replicas were required to capture the broad ensemble of structures in the denatured state.[75] This study of ACBP was extended by using an all-atom representation of the protein, thereby incorporating structural information from a molecular mechanics force field to supplement the PRE distance restraints.[78] A similar approach was used to model the disordered ensemble of α-synuclein.[15] Distance restraints were combined with all-atom MD simulations. The unfolded ensemble was found to be more compact than a random coil and to contain no secondary structure. The long range contacts between the hydrophobic region and the C-terminus favour the collapsed state of α-synuclein, and may offer some protection against aggregation.[15] Importantly, the hydrodynamic radius obtained using the distance-restrained MD simulations was consistent with the experimentally-determined value, supporting the validity of the approach.[15]

Applying experimentally-derived restraints in MD simulations ensures that the simulations produce ensembles consistent with what is known experimentally. However, a few cautionary notes must be kept in mind when using experimental-theoretical hybrid approaches. First, the experimental methods used to obtain the restraints may perturb the ensembles they are characterizing. In order to obtain PRE measurements of distances, nitroxide spin labels are coupled to cysteine residues in the protein. Both the introduction of spin labels and the mutation of residues to cysteine may result in perturbations in the conformations in the ensemble, as well as their relative populations.[25] Using experimental restraints obtained under one set of thermodynamic conditions (e.g. temperature, pH, ionic strength and pressure) in an MD simulation with completely different conditions will produce an unphysical conformational ensemble.[69] The simulation system must capture the *in vitro* conditions as closely as possible. When comparing simulation data with experimental observables, the same considerations apply.[69] It is also recommended that only primary (directly observable) experimental data be used as restraints, if possible. Secondary data that is computed using approximations and

assumptions based on primary data, such as S^2 order parameters, may result in unexpected artefacts.[69] In order to validate the ensembles generated by hybrid approaches, experimental data not used as restraints in the initial calculation can be used.[14, 25] Cross-validation is also possible by using a subset of the restraints to generate the ensemble, then verifying that the remaining restraints are reproduced.[14] In principle, MD simulations do not require any experimental data as input, and therefore can be used independently to predict experimental measurements.

DE NOVO MD SIMULATIONS

Lessons from Simulations of Unfolded States

MD simulations of disordered proteins are complicated by the fact that not one but possibly many thousands of states must be sampled in order to characterize a structurally heterogeneous ensemble. Barriers to conformational transitions involving significant structural rearrangements are high, and consequently transitions between conformations in the ensemble are statistically rare events. This is an important issue because simulation times are limited by current computer technology. Similar challenges are encountered in simulations of the unfolded state or partially-unfolded state of globular proteins.[79] There are similarities between disordered ensembles and unfolded or denatured ensembles.[25] It is therefore instructive to consider the extensive work on simulations of unfolded states and denatured states of proteins with folded states. In order to accurately characterize an ensemble, statistical convergence must be reached, and the model used to represent the polypeptide must be sufficiently accurate. However, accurate simulations are generally also computationally demanding.

One approach to this statistical sampling challenge is to run thousands of independent MD simulations in parallel for tens of nanoseconds each using a distributed computing 'supercluster'.[79, 80] In this approach, many conformations of the unfolded ensemble are simulated simultaneously starting from the fully extended state, thereby overcoming the problem of conformational transitions. The main drawback of this approach is the significant computational expense, requiring microseconds of simulation.[79] Simulations of the unfolded ensemble of villin headpiece, the tryptophan zipper and BBA5 resulted in the formulation of the "mean structure hypothesis". The unfolded ensemble has the same mean structure as the folded state, only with a much higher structural diversity. The average distance matrix of Cα pairs of the unfolded ensemble is very similar to that of the folded state for all three proteins.[79] The results of this study underscore the utility of studying average properties of disordered ensembles. Another approach to studying unfolded ensembles has

been the use of biased molecular dynamics (BMD) to preferentially sample conformations with different radii of gyration.[81] Starting from the native state of α-lactalbumin, a biasing force is used to generate conformers of larger radii of generation, which are subsequently simulated using unbiased MD in order to model the partially-denatured, molten globule-like state.[81]

MD Simulations of IDPs

MD simulations of disordered proteins in atomistic detail with an explicit representation of the solvent are computationally expensive. They are currently limited to the nanosecond to microsecond timescale for continuous simulations depending on the size of the system and the available computational resources. For example, the disordered protein α-synuclein was recently studied using MD simulations as a monomer in solution and as an aggregate with a membrane (with more than 200,000 atoms) for several nanoseconds.[82, 83] It is important to consider statistical convergence in addition to atomistic detail. Without achieving convergence, it is not possible to draw meaningful conclusions from the simulations.[84] The two most straightforward approaches involve running many short simulations (MS) or sampling a few long trajectories (FL).[85] These two approaches were rigorously compared with explicit solvent simulations of the RN24 peptide (totalling more than 800 μs).[85] Although structural properties obtained with both FL and MS simulations agreed qualitatively, MS simulations resulted in greater precision. Short timescale conformational transitions observed in MS simulations were not observed in the FL simulations. Conversely, a few transitions occurring over long timescales were observed in the FL simulations and not the MS simulations. The FL and MS approaches are therefore complementary.[85] An approach intermediate between MS and FL is also possible (i.e. several simulations of intermediate length). For example, the intrinsically disordered transactivation of p53 was simulated in explicit solvent for six simulations totalling nearly 0.5 μs. The resulting conformational ensemble was dominated by compact states with significant amounts of secondary structure, largely due to the high composition of leucine.[86] MS simulations can also involve more complex implementations, including Markov State Models.[87]

Energy landscapes of biomolecules are “rugged” and the source of this ruggedness is two-fold. The energetic barriers separating accessible states are often larger than the available thermal energy, and there are typically a large number of states to be sampled. Consequently, conventional or “brute force” MD simulations are often insufficient to achieve complete Boltzmann sampling of all important conformational states. For this reason, generalized-ensemble algorithms utilizing a random walk in temperature have become popular tools for conformational sampling. These methods rely on the fact that the free energy surface becomes less rugged at high temperature, increasing the frequency of interconversion between conformational states.[88] Replica

exchange is the most commonly used method to enhance sampling in biomolecular simulation.[89-91] It has been used to study the conformational ensemble of polyalanine. The relative populations of α-helical, polyproline II and disordered conformations were quantified as a function of temperature.[92] Refer to reference 93 for a review and thorough comparison of generalized-ensemble algorithms for conformational sampling.[93]

Simulations of Non-folders

We recently developed and tested a novel generalized ensemble algorithm, simulated tempering distributed replica sampling (STDR),[93] which is a combination of simulated tempering [94, 95] and distributed replica sampling.[96-98] STDR makes use of a random walk in temperature to enhance conformational sampling. Unlike replica exchange, STDR does not require a synchronous cluster of computers and is well-suited to distributing computing platforms. It is more computationally efficient than simulated tempering for complex biomolecular systems because it requires less initial simulation to achieve homogeneous temperature sampling. STDR was used to study the elastin-like peptide $(GVPGV)_7$ in atomic detail with explicit water, for a total simulation time of more than 8 μs. Representative structures from the structurally heterogeneous ensemble are shown with the distribution of the radius of gyration n Figure 2. Compared to MD simulations in the canonical ensemble, STDR dramatically enhances conformational sampling while offering significant practical advantages compared to replica exchange and simulated tempering.[93]

Pappu and co-workers have extensively characterized the conformational ensemble of polyglutamine, which is implicated in Huntingtin's disease.[99-102] Using MD simulations, polyglutamine was shown shown to adopt a heterogeneous ensemble of collapsed structures. Polyglutamine is disordered as a monomer because the glutamine side chains form hydrogen bonds to the backbone, essentially competing with water as a solvent.[99] Individual residues have significant populations in the PPII and α-helical regions of the Ramachandran plot, but there is no extended PPII helix or α-helix. The result is a conformational entropy 'bottleneck' that must be overcome in order for aggregation and β-sheet formation to occur. While statistical fluctuations lead to transient sampling of β-hairpin-like states,[99] conformations with a high content of β-type structures are thermodynamically-disfavoured for polyglutamine monomers.[101]

The results from the MD simulations are consistent with the hydrodynamic properties of polyglutamine measured using fluorescence correlation spectroscopy. Chain size, as measured by the radius of hydration, is related to chain length by a power law. Polyglutamine in aqueous solution behaves like a polymer in a poor solvent, for which chain-chain interactions are preferred to chain-solvent interactions.[100] This behaviour is not

unique to polyglutamine. In fact, water is also a poor solvent for polyglycine and poly(glycine-serine), both of which cannot make sidechain-backbone hydrogen bonds.[103] In 8M urea, both polyglycine and poly(glycine-serine) have conformational ensembles that are more swollen than the collapsed ensembles in water.[103] Despite the lack of a hydrophobic core, polar but uncharged sequences collapse in water suggesting that water is a poor solvent for the polypeptide backbone.[6, 100, 103] These observations have direct implications for the aggregation of polyglutamine, since phase separation (aggregation) only occurs in a poor solvent, and increasing chain length increases the driving force for aggregation.[104]

Another class of non-folders are the nucleoporins, which function as entropic chains. Selective transport between the nucleus and cytoplasm occurs through the nuclear pore complex (NPC). The size-dependent selectivity of the NPC is due to the phenylalanine-glycine nucleoporins (FG-nups), which are intrinsically disordered.[105] FG-nups coating a nanopore of the same dimensions as the NPC are sufficient to reproduce the NPC's selectivity.[106] An FG domain (of 111 residues), as well as a mutant with all phenylalanines substituted with alanines, were studied using both MD in implicit solvent and NMR. Self diffusion coefficients of both the FG domain and the F→A mutant were measured using NMR. Both simulation and experiment characterized the ensemble of the FG domain to be significantly more compact than the F→A mutant. The conformational ensemble of the FG domain is best described as a native pre-molten globule, while the F→A mutant is more coil-like.[105] Using spectral clustering, conformational sampling of many short MD simulations of FG-nups was assessed in detail.[107] Clustering is useful both for rigorously defining conformational states,[17, 108] as well as offering a means to evaluate the extent of conformational sampling.[109, 110]

Average properties of disordered ensembles, such as interatomic distances, provide useful information that can be compared with experimental data that is also ensemble-averaged.[79] Average properties of an ensemble may not be apparent when looking at any single structure.[79] However, sampling a small subset of the conformations from a disordered ensemble may be sufficient for convergence of average properties. We observed this convergence in our study of disordered elastin-like and amyloid-like peptides.[38] Since the simulations were relatively short (60 ns for monomers and 20 ns for aggregates) and no enhanced sampling approach was used, the peptides explored only a small subset of their complete conformational ensembles. Despite these limitations, convergence of the average number of backbone-backbone hydrogen bonds, backbone-water hydrogen bonds and PPII content was observed. Elastin-like peptides have fewer intra-peptide hydrogen bonds and are more hydrated on average compared to amyloid-like peptides.[38]

Simulations of Disordered Protein-Protein Interactions

MD simulations can be used to complement structural information from NMR spectroscopy in the study of binding complexes of IDPs. This approach is particularly useful if the ligand remains partially disordered in the bound state. The interaction of a 15-residue peptide from mouse guanine-nucleotide exchange factor Sos2 and the SH3 (Src-homology-3) domain of mouse Grb2 was studied using MD simulation.[111] An NMR structure was obtained for the complex, but complete assignment of the peptide resonances was not possible. Consequently, it could not be concluded if the peptide adopts a PPII helix conformation, similar to the bound state of other SH3 ligands.[112] NMR provided sufficient data to determine the orientation of the peptide and the location of only two residues. The binding site of the SH3 domain is a hydrophobic patch on the surface of the protein. Starting from this limited data, MD simulations provided a conformational ensemble describing the bound state of the peptide. The peptide was found to adopt a predominantly extended conformation in contact with the SH3 domain, stabilized by sidechain-sidechain and sidechain-backbone hydrogen bonds. The peptide retains significant flexibility in the bound state, and the ideal PPII helix proposed in the NMR study as the structure of the peptide is not compatible with the conformational ensemble of the MD simulation.[111]

Monte Carlo simulations were used to simulate the $p27^{Kip1}$ protein bound to the cyclin A-cyclin-dependent kinase 2 (Cdk2) complex.[29] The $p27^{Kip1}$ protein is intrinsically disordered in its unbound state, and orders upon binding to the cyclin A-Cdk2 complex. Simulations were initiated from the crystal structure of the bound state in order to model the transition state ensemble, with the cyclin A-Cdk2 complex held fixed. A large ensemble of unfolding-unbinding trajectories were generated at high temperature. A significant amount of native-like topology was found in the transition state. It is, however, problematic to infer information about the binding free energy landscape at physiological temperature from simulations at high temperature.[113] Subsequent simulations of the order-disorder transition of the $p27^{Kip1}$ protein with the cyclin A-Cdk2 complex were performed using simulated annealing.[113] When $p27^{Kip1}$ is bound to the cyclin A-Cdk2 complex, its C-terminal region remains disordered. The conformational ensemble of the C-terminal region was characterized using both MD simulations and SAXS measurements to be extended and flexible in the bound state.[114]

Simulation and Experiment – Interdisciplinary Studies

MD simulations are readily combined with other biophysical approaches. Simulation studies are (inherently) single molecule studies. Since biomolecular systems are ergodic, ensemble averages can be calculated from trajectories and compared to experimental measurements. In order to determine how to most

appropriately simulate disordered states of proteins, it is essential to validate theoretical approaches using appropriate experimental data. One example of such a rigorous validation is a recent study in which the conformational ensembles of Aβ40 and Aβ42 were studied using replica exchange MD in explicit water on the microsecond timescale. A comparison was made between different force fields for the Aβ42 peptide. Scalar couplings determined from NMR experiments were in quantitative agreement with the results of the simulations using the OPLS force field. This work validated the conformational ensembles generated using replica exchange in addition to providing a ranking of force fields.[115]

In another interdisciplinary study, different solvent representations were compared using a set of unstructured peptides as test systems.[116] This comparison was validated using experimental data from triplet-triplet energy transfer (TTET) experiments.[116] TTET can be used to monitor loop formation in unfolded polypeptides.[117] Since contact formation between residues occurs on a timescale accessible to MD simulation, a direct comparison of the kinetics obtained using simulation and experiment was performed.[116] Using this interdisciplinary approach, polyserine and poly(glycine-serine) peptides were found to have little persistent structure and non-Gaussian end-to-end distance distributions, indicative of rugged energy landscapes.[116] The use of an explicit solvent (TIP4P) resulted in better agreement with TTET experiments compared to implicit solvent representations, which overestimated compactness, secondary structure content and the number of peptide-peptide hydrogen bonds.[116] While implicit solvent provides advantages in terms of calculation speed, the compromise in accuracy of the description of the conformational ensemble of IDPs may be significant.

Replica exchange simulations of the GB1 peptide with an implicit solvent were compared with NMR experiments. Chemical shifts and scalar couplings were calculated for the ensembles at each temperature in the replica exchange simulation. The ensemble of structures that most closely matched the NMR data obtained at 278 K corresponded to temperatures near 400 K.[118] Comparisons with both NMR and fluorescence experiments suggest that implicit solvent models result in overly-structured ensembles of IDPs at low temperature. A promising new implicit solvent model, ABSINTH (self-assembly of biomolecules studied by an implicit, novel and tunable Hamiltonian), was recently developed. It is calibrated primarily for simulating the conformational ensembles of IDPs.[119] ABSINTH is designed to maximize computational efficiency, requiring only 2.5 to 5 times the expense of simulations in vacuo. In order to better understand how sequence dictates a disordered conformational ensemble, high-throughput studies of many IDPs are required. Reaching statistical convergence for such ensembles is computationally intensive, and ABSINTH is well-suited for such high-throughput MD studies.[119] Implicit solvent representations represent a significant simplification compared to explicit solvent. Similarly, coarse-grained simulations involve simplified representations of the polypeptide chain.

Coarse-Grained Simulations

Simulations of proteins can include different levels of detail, ranging from lattice models to all-atom representations. So-called 'minimalist models' offer several advantages, and have already been very useful in studies of protein folding. First, the computational cost is dramatically reduced, thereby facilitating simulations of significantly larger systems and/or longer timescales. Comparing simulations of coarse-grained models to experiment also offers the possibility of determining thespecific properties of proteins that result in observed experimental behaviours.[120] Even if the model does not reproduce the experimental data, it is possible to learn from the result and refine the model. Validation of minimalist models with experiment is essential to determine if the elements of the model are sufficient.[120]

A Cα-model was recently developed specifically for the simulation of unfolded and disordered ensembles.[121] The energy parameters were optimized through an iterative process incorporating PRE measurements of the unfolded ensemble of the Δ131Δ fragment of Staphylococcal nuclease. An initial guess of the energy parameters was used to generate an unfolded ensemble, for which PRE measurements were back-calculated. An optimization scheme was used to obtain new energy parameters in better agreement with the experimental data, and the process was iterated. In order to obtain transferable energy parameters, PRE measurements from other disordered ensembles are required and other types of experimental data can be readily incorporated.[121]

One of the simplest models used to represent a polypeptide chain is the worm-like chain (WLC) model.[122] The polypeptide is modeled as a continuous cylinder with a randomly-directed radius of curvature using only two parameters, contour length, l_c, and persistence length, l_p. For a polypeptide, l_c is the number of residues multiplied by the distance per residue (3.8 Å) and l_p is the chain length after which the direction of the tangent vector to the cylinder is uncorrelated.[123] Flexible linkers connecting independently folded protein domains have successfully been modeled as worm-like chains with an l_p of 4.0 Å,[123, 124] while shorter loops were modeled with an l_p of 3.04 Å.[123, 125] (Refer to reference 123 for a review of the use of polymer models to represent unfolded states and flexible linkers.[123])

Simulations of disordered proteins have helped to elucidate the biophysical underpinnings of the sequence characteristics uncovered by bioinformatics.[48, 126] For example, it is known that the compositions of disordered sequences are depleted in hydrophobic residues and enhanced in charged residues.[21] Based on a small dataset of natively-unfolded and folded proteins, it was suggested that a combined criteria of net charge

and hydrophobicity defines a border between disordered and ordered sequences.[46] However, it was shown using a larger dataset that the border in the charge-hydrophobicity plot is not well-defined, with significant overlap between ordered and disordered sequences (similar to Figure 1b).[48] This "twilight zone" between order and disorder was investigated using a 2d-lattice model. The polypeptide chain was represented using an HP model with hydrophobic and polar residues, and an HPN model with hydrophobic, positive and negative residues. Conformations of peptides of varying length and composition were generated. Remarkably, it was found using this simple model that specific interactions are more important in short chains, in agreement with the bioinformatics analysis of the compositions of disordered and ordered sequences.[48] Sequence composition was found to be a better determinant of disorder for longer sequences. Importantly, the results of this study suggest that disorder predictors based on composition alone may never be sufficiently accurate for short peptides.[48] In a related coarse-grained simulation study, Ashbaugh and Hatch showed that there is a coil-to-globule collapse driven by a combination of charge and hydropathy.[126]

A possible role for disordered regions flanking linear binding motifs was also recently elucidated using coarse-grained simulations.[127] In Abeln and Frenkel's model of a polypeptide, each residue occupies a point on a cubic lattice and interacts with other residues via a pairwise interaction energy, using a Monte Carlo algorithm for conformational sampling. Remarkably, using this simple representation, it was shown that flexible hydrophobic binding motifs that are not flanked by disordered regions are prone to aggregation. Without the disordered flanks, hydrophobic linear motifs are suggested to be toxic. The function of the disordered regions flanking the binding motif is to impede aggregate formation without obstructing substrate binding.[127]

CONCLUSION

This review is a summary of the early work on simulations of disordered states, as well as what is currently state-of-the-art. MD simulations have already provided structural biologists with an unprecedented view of the conformational ensembles of disordered states. However, the application of molecular simulations to study disordered proteins is still in its early stages. The capability of high performance computing is expected to improve, facilitating simulations of longer timescales and larger systems,[128] including all-atom detail and explicit solvent. The development of enhanced sampling algorithms has made it possible to achieve statistical convergence for disordered state ensembles. As evidenced by many studies reviewed here, the dialogue between experimentalists and theorists is particularly valuable for elucidating the complexities of disordered states. MD simulations are well-suited to make use of experimental restraints including PRE measurements, chemical shifts and SAXS data. Experimental data, usually in the form of ensemble averages, is insufficient to

completely determine a conformational ensemble and can be supplemented with the information in force fields. Unrestrained MD simulations have also been used to characterize disordered states. Simulations are therefore useful for making predictions, especially for NMR spectroscopy, SAXS and fluorescence studies. In turn, these studies provide an important validation of the force fields and an understanding of their limitations. In future, the field would benefit from more of these fruitful interdisciplinary studies. In addition to providing populations of different conformations in the ensembles of disordered states, simulations can also provide information on the dynamics of single molecules. Interconversion between different conformations can be directly observed in atomistic detail and lifetimes of conformations can be determined. Simulations can provide a complete characterization of the individual conformations in a disordered ensemble, information not currently accessible by experiment. MD simulations therefore represent an important, but so far under-utilized tool, for the study of protein disorder.

FIGURES

A **B**

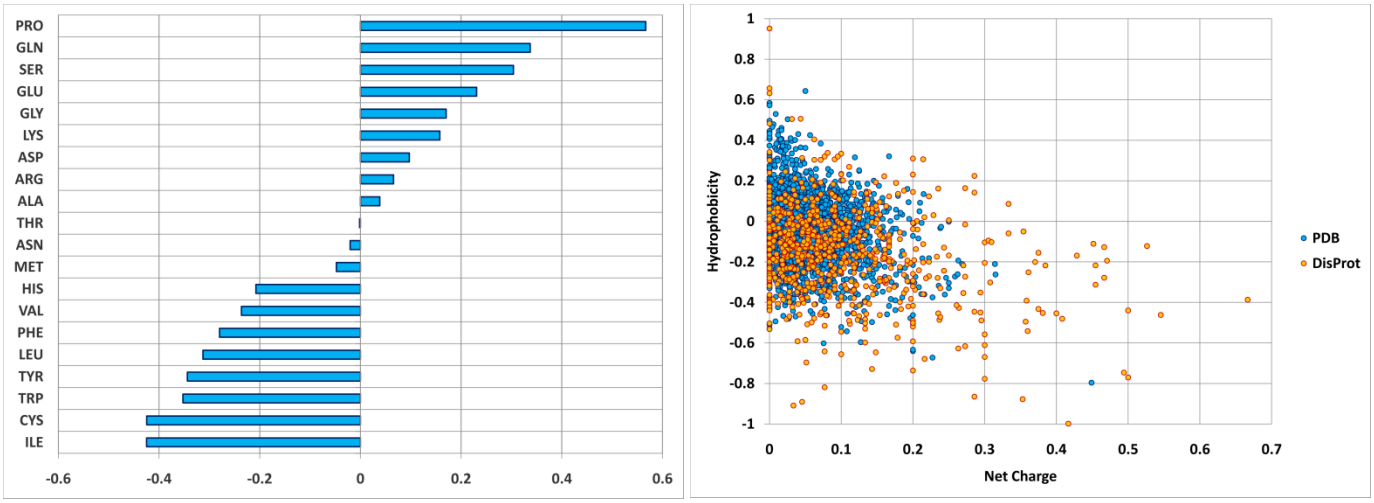
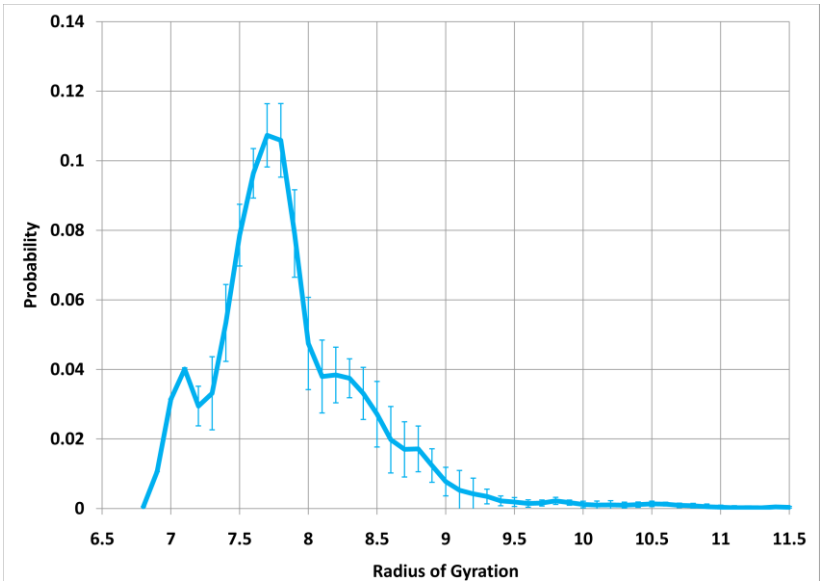


Figure 1: Sequence Features of Disordered Proteins

A set of proteins from the PDB with less than 20 % sequence identity was obtained using the PISCES server.[129] This non-redundant set represents ordered proteins. They are known to be structured by virtue of the fact that their structures have been solved by NMR or x-ray crystallography. The set of disordered regions was obtained from the current release of DISPROT (version 4.9). Only regions with more than five contiguous residues were included in the data set. Amino acid composition was computed individually for each sequence and averaged over the data set. In (A), the average composition of DISPROT minus the average composition of the PDB, normalized by the average composition of the PDB [$(C_{DISPROT} - C_{PDB})/ C_{PDB}$] is shown. The amino acids that are enriched in DISPROT relative to the PDB are P, Q, S, E, G, K, D, R and A, while T, N, M, H, V, F, L, Y, W, C and I are depleted. In (B), net charge is plotted versus hydrophobicity. Hydrophobicity was computed using the normalized scale of Sweet and Eisenberg.[130] Net charge was computed by taking aspartic acid and glutamic acid to have a charge of -1 and lysine and arginine to have a charge of +1, then normalizing by the chain length.

A



B

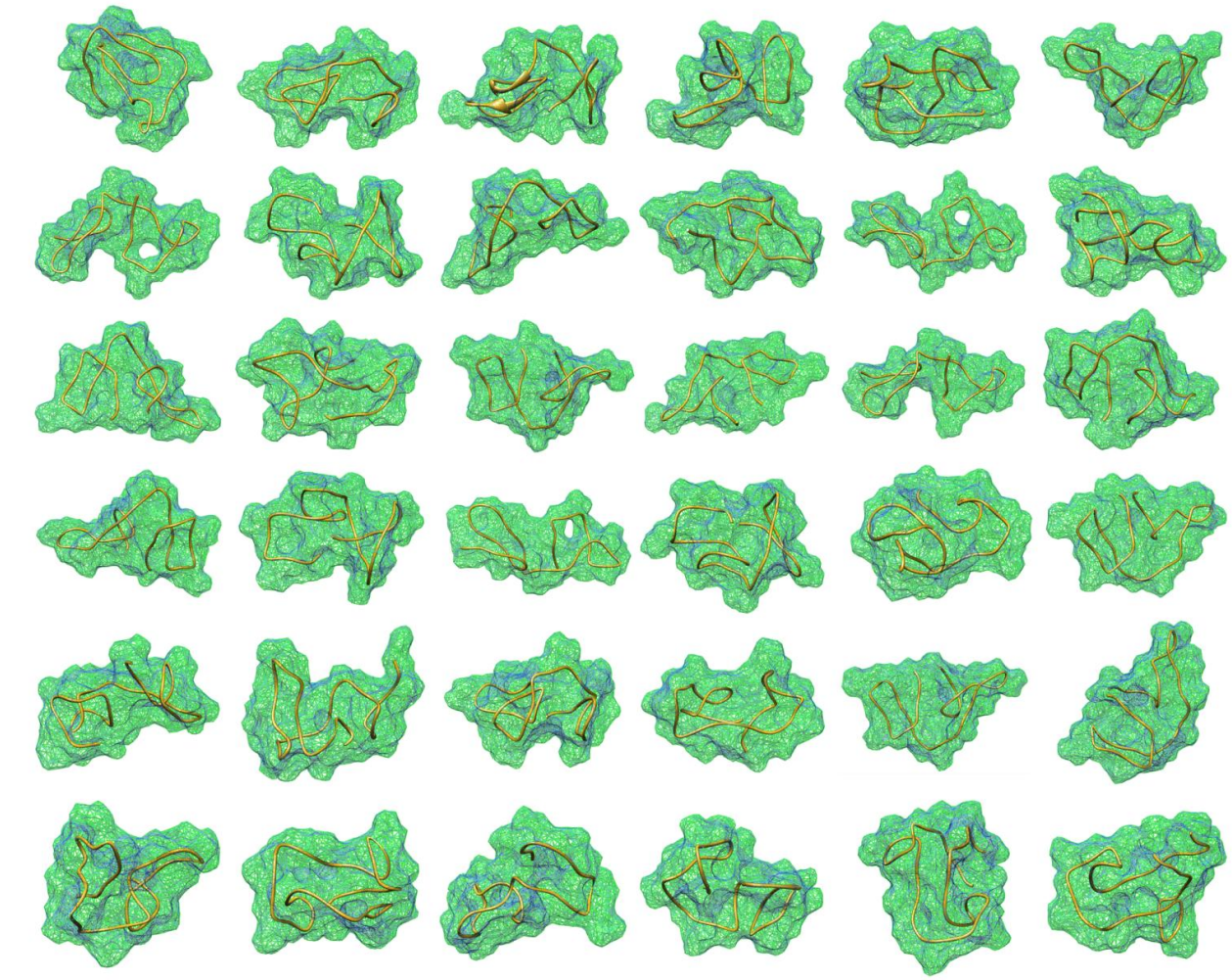


Figure 2: The Disordered Ensemble of $(GVPGV)_7$

(A) Radius of gyration distribution and (B) representative structures[131] from the conformational ensemble.

REFERENCES

1. Wright, P. E.; Dyson, H. J., *J. Mol. Biol.* **1999,** *293* (2), 321-331.
2. Fischer, E., *Ber. Deutsch. Chem. Ges.* **1894,** *27*, 2985.
3. Anfinsen, C. B., *Science* **1973,** *181* (4096), 223-230.
4. Mirsky, A. E.; Pauling, L., *Proc. Natl. Acad. Sci. U. S. A.* **1936,** *22*, 439-447.
5. Eliezer, D., *Curr. Opin. Struct. Biol.* **2009,** *19* (1), 23-30.
6. Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L., *Curr. Opin. Struct. Biol.* **2008,** *18* (6), 756-764.
7. Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T., *J. Mol. Biol.* **2004,** *337* (3), 635-645.
8. Tompa, P., *FEBS Lett.* **2005,** *579* (15), 3346-3354.
9. Uversky, V. N., *Protein Sci.* **2002,** *11* (4), 739-756.
10. Bracken, C.; Iakoucheva, L. M.; Rorner, P. R.; Dunker, A. K., *Curr. Opin. Struct. Biol.* **2004,** *14* (5), 570-576.
11. Uversky, V. N.; Dunker, A. K., *Science* **2008,** *322* (5906), 1340-1341.
12. Uversky, V. N.; Oldfield, C. J.; Dunker, A. K., *Ann. Rev. Biophys.* **2008,** *37*, 215-246.
13. Dunker, A. K.; Oldfield, C. J.; Meng, J. W.; Romero, P.; Yang, J. Y.; Chen, J. W.; Vacic, V.; Obradovic, Z.; Uversky, V. N., *BMC Genomics* **2007,** *9*, 25.
14. Vendruscolo, M., *Curr. Opin. Struct. Biol.* **2007,** *17* (1), 15-20.
15. Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M., *J. Am. Chem. Soc.* **2005,** *127* (2), 476-477.
16. Dobson, C. M.; Sali, A.; Karplus, M., *Angew. Chem.-Int. Edit.* **1998,** *37* (7), 868-893.
17. Daura, X., *Theor. Chem. Acc.* **2006,** *116* (1-3), 297-306.
18. Daura, X.; Glattli, A.; Gee, P.; Peter, C.; Van Gunsteren, W. F., Unfolded state of peptides. In *Unfolded Proteins*, Academic Press Inc: San Diego, 2002; Vol. 62, pp 341-360.
19. Dill, K. A.; Shortle, D., *Annu. Rev. Biochem.* **1991,** *60*, 795-825.
20. Dyson, H. J.; Wright, P. E., *Chem. Rev.* **2004,** *104* (8), 3607-3622.
21. Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. R.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C. H.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, M.; Garner, E. C.; Obradovic, Z., *Journal of Molecular Graphics and Modelling* **2001,** *19* (1), 26-59.
22. Yang, D. W.; Mittermaier, A.; Mok, Y. K.; Kay, L. E., *J. Mol. Biol.* **1998,** *276* (5), 939-954.
23. Baldwin, R. L.; Zimm, B. H., *Proc. Natl. Acad. Sci. U. S. A.* **2000,** *97* (23), 12391-12392.
24. Pappu, R. V.; Srinivasan, R.; Rose, G. D., *Proc. Natl. Acad. Sci. U. S. A.* **2000,** *97* (23), 12565-12570.
25. Mittag, T.; Forman-Kay, J. D., *Curr. Opin. Struct. Biol.* **2007,** *17* (1), 3-14.
26. Dyson, H. J.; Wright, P. E., *Nat. Rev. Mol. Cell Biol.* **2005,** *6* (3), 197-208.
27. Xue, B.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N., *FEBS Lett.* **2009,** *583* (9), 1469-1474.
28. Wright, P. E.; Dyson, H. J., *Curr. Opin. Struct. Biol.* **2009,** *19* (1), 31-38.
29. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. W., *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100* (9), 5148-5153.
30. Tompa, P., *Trends Biochem.Sci.* **2002,** *27* (10), 527-533.
31. Tompa, P.; Fuxreiter, M., *Trends Biochem.Sci.* **2008,** *33* (1), 2-8.
32. Borg, M.; Mittag, T.; Pawson, T.; Tyers, M.; Forman-Kay, J. D.; Chan, H. S., *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (23), 9650-9655.
33. Mittag, T.; Orlicky, S.; Choy, W. Y.; Tang, X. J.; Lin, H.; Sicheri, F.; Kay, L. E.; Tyers, M.; Forman-Kay, J. D., *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105* (46), 17772-17777.
34. Brocca, S.; Samalikova, M.; Uversky, V. N.; Lotti, M.; Vanoni, M.; Alberghina, L.; Grandori, R., *Proteins* **2009,** *76* (3), 731-746.
35. Mohan, A.; Sullivan, W. J.; Radivojac, P.; Dunker, A. K.; Uversky, V. N., *Mol. Biosyst.* **2008,** *4* (4), 328-340.
36. Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z., *Biochemistry* **2002,** *41* (21), 6573-6582.
37. Hoh, J. H., *Proteins* **1998,** *32* (2), 223-228.

38. Rauscher, S.; Baud, S.; Miao, M.; Keeley, F. W.; Pomès, R., *Structure* **2006,** *14* (11), 1667-1676.
39. Cao, Y.; Li, H. B., *Nat. Nanotechnol.* **2008,** *3* (8), 512-516.
40. Gsponer, J.; Futschik, M. E.; Teichmann, S. A.; Babu, M. M., *Science* **2008,** *322* (5906), 1365-1368.
41. Song, J. X., *FEBS Lett.* **2009,** *583* (6), 953-959.
42. Uversky, V. N.; Fink, A. L., *BBA-Proteins Proteomics* **2004,** *1698* (2), 131-153.
43. Calamai, M.; Chiti, F.; Dobson, C. M., *Biophys. J.* **2005,** *89* (6), 4201-4210.
44. Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I., *J. Mol. Biol.* **2005,** *347* (4), 827-839.
45. Romero, P.; Obradovic, Z.; Li, X. H.; Garner, E. C.; Brown, C. J.; Dunker, A. K., *Proteins-Structure Function and Genetics* **2001,** *42* (1), 38-48.
46. Uversky, V. N.; Gillespie, J. R.; Fink, A. L., *Proteins-Structure Function and Genetics* **2000,** *41* (3), 415-427.
47. Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N.; Obradovic, Z.; Dunker, A. K., *Nucleic Acids Res.* **2007,** *35*, D786-D793.
48. Szilagyi, A.; Gyorffy, D.; Zavodszky, P., *Biophys. J.* **2008,** *95* (4), 1612-1626.
49. Williams, R. M.; Obradovic, Z.; Mathura, V.; Braun, W.; Garner, E. C.; Young, J.; Takayama, S.; Brown, C. J.; Dunker, A. K., *Pac. Symp. Biocomput.* **2001,** *6*, 89-100.
50. Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H., *FEBS Lett.* **2004,** *576* (3), 348-352.
51. Wootton, J. C., *Computers & Chemistry* **1993,** *18*, 269-285.
52. Barrick, D.; Ferreiro, D. U.; Komives, E. A., *Curr. Opin. Struct. Biol.* **2008,** *18* (1), 27-34.
53. Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I., *Bioinformatics* **2005,** *21* (16), 3433-3434.
54. Radivojac, P.; Iakoucheva, L. M.; Oldfield, C. J.; Obradovic, Z.; Uversky, V. N.; Dunker, A. K., *Biophys. J.* **2007,** *92* (5), 1439-1456.
55. Ferron, F.; Longhi, S.; Canard, B.; Karlin, D., *Proteins* **2006,** *65* (1), 1-14.
56. Seema Sharma; Haiyan Zheng; Yuanpeng J. Huang; Asli Ertekin; Yoshitomo Hamuro; Paolo Rossi; Roberto Tejero; Thomas B. Acton; Rong Xiao; Mei Jiang; Li Zhao; Li-Chung Ma; G. V. T. Swapna; James M. Aramini; Gaetano T. Montelione, *Proteins: Structure, Function, and Bioinformatics* **2009,** *76* (4), 882-894.
57. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K., *Biochemistry* **2005,** *44* (6), 1989-2000.
58. Marsh, J. A.; Singh, V. K.; Jia, Z. C.; Forman-Kay, J. D., *Protein Sci.* **2006,** *15* (12), 2795-2804.
59. Receveur-Brechot, V.; Bourhis, J. M.; Uversky, V. N.; Canard, B.; Longhi, S., *Proteins* **2006,** *62* (1), 24-45.
60. Oh, K. J.; Cash, K. J.; Plaxco, K. W., *Chem.-Eur. J.* **2009,** *15* (10), 2244-2251.
61. Gustiananda, M.; Liggins, J. R.; Cummins, P. L.; Gready, J. E., *Biophys. J.* **2004,** *86* (4), 2467-2483.
62. Choy, W. Y.; Forman-Kay, J. D., *J. Mol. Biol.* **2001,** *308* (5), 1011-1032.
63. Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W. Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D., *J. Mol. Biol.* **2007,** *367* (5), 1494-1510.
64. Marsh, J. A.; Forman-Kay, J. D., *J. Mol. Biol.* **2009,** *391* (2), 359-374.
65. Feldman, H. J.; Hogue, C. W. V., *Proteins* **2000,** *39* (2), 112-131.
66. Bezsonova, I.; Evanics, F.; Marsh, J. A.; Forman-Kay, J. D.; Prosser, R. S., *J. Am. Chem. Soc.* **2007,** *129* (6), 1826-1835.
67. Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I., *J. Am. Chem. Soc.* **2007,** *129* (17), 5656-5664.
68. Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M., *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102* (47), 17002-17007.
69. van Gunsteren, W. F.; Dolenc, J.; Mark, A. E., *Curr. Opin. Struct. Biol.* **2008,** *18* (2), 149-153.
70. Fuentes, G.; Nederveen, A. J.; Kaptein, R.; Boelens, R.; Bonvin, A. M. J. J., *J. Biomol. NMR* **2005,** *33* (3), 175-186.
71. Huang, A.; Stultz, C. M., *PLoS Comput. Biol.* **2008,** *4* (8), 12.
72. Yoon, M. K.; Venkatachalam, V.; Huang, A.; Choi, B. S.; Stultz, C. M.; Chou, J. J., *Protein Sci.* **2009,** *18* (2), 337-347.
73. Torda, A. E.; Scheek, R. M.; van Gunsteren, W. F., *Chem. Phys. Lett.* **1989,** *157* (4), 289-294.
74. Gros, P.; van Gunsteren, W. F.; Hol, W. G. J., *Science* **1990,** *249* (4973), 1149-1152.

75. Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M., *J. Am. Chem. Soc.* **2004,** *126* (10), 3291-3299.
76. Makowska, J.; Rodziewicz-Motowidlo, S.; Baginska, K.; Vila, J. A.; Liwo, A.; Chmurzynski, L.; Scheraga, H. A., *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (6), 1744-1749.
77. Christen, M.; Keller, B.; van Gunsteren, W. F., *J. Biomol. NMR* **2007,** *39* (4), 265-273.
78. Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M., *J. Mol. Biol.* **2005,** *347* (5), 1053-1062.
79. Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S., *J. Mol. Biol.* **2002,** *323* (1), 153-164.
80. Zagrovic, B.; Pande, V. S., *J. Am. Chem. Soc.* **2006,** *128* (36), 11742-11743.
81. Paci, E.; Smith, L. J.; Dobson, C. M.; Karplus, M., *J. Mol. Biol.* **2001,** *306* (2), 329-347.
82. Tsigelny, I. F.; Sharikov, Y.; Miller, M. A.; Masliah, E., *Nanomed.-Nanotechnol. Biol. Med.* **2008,** *4* (4), 350-357.
83. Tsigelny, I. F.; Bar-On, P.; Sharikov, Y.; Crews, L.; Hashimoto, M.; Miller, M. A.; Keller, S. H.; Platoshyn, O.; Yuan, J. X. J.; Masliah, E., *Febs J.* **2007,** *274* (7), 1862-1877.
84. Lyman, E.; Zuckerman, D. M., *Biophys. J.* **2006,** *91* (1), 164-172.
85. Monticelli, L.; Sorin, E. J.; Tieleman, D. P.; Pande, V. S.; Colombo, G., *J. Comput. Chem.* **2008,** *29* (11), 1740-1752.
86. Espinoza-Fonseca, L. M., *FEBS Lett.* **2009,** *583* (3), 556-560.
87. Singhal, N.; Snow, C. D.; Pande, V. S., *J. Chem. Phys.* **2004,** *121* (1), 415-425.
88. Huang, X.; Bowman, G. R.; Pande, V. S., *J. Chem. Phys.* **2008,** *128* (20), 15.
89. Sugita, Y.; Okamoto, Y., *Chem. Phys. Lett.* **1999,** *314* (1-2), 141-151.
90. Tesi, M. C.; van Rensburg, E. J. J.; Orlandini, E.; Whittington, S. G., *J. Stat. Phys.* **1996,** *82* (1-2), 155-181.
91. Ferrenberg, A. M.; Swendsen, R. H., *Phys. Rev. Lett.* **1988,** *61* (23), 2635-2638.
92. Garcia, A. E., *Polymer* **2004,** *45* (2), 669-676.
93. Rauscher, S.; Neale, C.; Pomès, R., *J. Chem. Theory Comput.* **2009,** *under review*.
94. Marinari, E.; Parisi, G., *Europhys. Lett.* **1992,** *19* (6), 451-458.
95. Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N., *J. Chem. Phys.* **1992,** *96* (3), 1776-1783.
96. Rodinger, T.; Howell, P. L.; Pomès, R., *J. Chem. Theory Comput.* **2006,** *2* (3), 725-731.
97. Rodinger, T.; Howell, P. L.; Pomès, R., *J. Chem. Phys.* **2008,** *129* (15), 12.
98. Neale, C.; Rodinger, T.; Pomès, R., *Chem. Phys. Lett.* **2008,** *460* (1-3), 375-381.
99. Wang, X. L.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V., *Proteins* **2006,** *63* (2), 297-311.
100. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V., *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (45), 16764-16769.
101. Vitalis, A.; Lyle, N.; Pappu, R. V., *Biophys. J.* **2009,** *97* (1), 303-311.
102. Vitalis, A.; Wang, X. L.; Pappu, R. V., *Biophys. J.* **2007,** *93* (6), 1923-1937.
103. Tran, H. T.; Mao, A.; Pappu, R. V., *J. Am. Chem. Soc.* **2008,** *130* (23), 7380-7392.
104. Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L., *Arch. Biochem. Biophys.* **2008,** *469* (1), 132-141.
105. Krishnan, V. V.; Lau, E. Y.; Yamada, J.; Denning, D. P.; Patel, S. S.; Colvin, M. E.; Rexach, M. F., *PLoS Comput. Biol.* **2008,** *4* (8), 13.
106. Jovanovic-Talisman, T.; Tetenbaum-Novatt, J.; McKenney, A. S.; Zilman, A.; Peters, R.; Rout, M. P.; Chait, B. T., *Nature* **2009,** *457* (7232), 1023-1027.
107. Phillips, J. L.; Colvin, M. E.; Lau, E. Y.; Newsam, S. In *Analyzing Dynamical Simulations of Intrinsically Disordered Proteins Using Spectral Clustering*, IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, PA, Nov 03-05; Chen, Y.; He, J.; Reddy, C. K.; Yang, J.; Yoo, I.; Zhang, X.; Gao, J.; Huang, Y.; Song, M.; Wu, Z., Eds. Ieee Computer Soc: Philadelphia, PA, 2008; pp 17-24.
108. Lowry, D. F.; Hausrath, A. C.; Daughdrill, G. W., *Proteins* **2008,** *73* (4), 918-928.
109. Frickenhaus, S.; Kannan, S.; Zacharias, M., *J. Comput. Chem.* **2009,** *30* (3), 479-492.
110. Smith, L. J.; Daura, X.; van Gunsteren, W. F., *Proteins* **2002,** *48* (3), 487-496.
111. Calero, S.; Lago, S.; van Gunsteren, W. F.; Daura, X., *Chem. Biodivers.* **2004,** *1* (3), 505-519.

112. Wittekind, M.; Mapelli, C.; Farmer, B. T.; Suen, K. L.; Goldfarb, V.; Tsao, J. L.; Lavoie, T.; Barbacid, M.; Meyers, C. A.; Mueller, L., *Biochemistry* **1994,** *33* (46), 13531-13539.
113. Verkhivker, G. M., *Proteins* **2005,** *58* (3), 706-716.
114. Galea, C. A.; Nourse, A.; Wang, Y.; Sivakolundu, S. G.; Heller, W. T.; Kriwacki, R. W., *J. Mol. Biol.* **2008,** *376* (3), 827-838.
115. Sgourakis, N. G.; Yan, Y. L.; McCallum, S. A.; Wang, C. Y.; Garcia, A. E., *J. Mol. Biol.* **2007,** *368* (5), 1448-1457.
116. Feige, M. J.; Paci, E., *J. Mol. Biol.* **2008,** *382* (2), 556-565.
117. Fierz, B.; Satzger, H.; Root, C.; Gilch, P.; Zinth, W.; Kiefhaber, T., *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (7), 2163-2168.
118. Weinstock, D. S.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R. M.; Wu, K. P.; Baum, J., *J. Am. Chem. Soc.* **2007,** *129* (16), 4858-+.
119. Vitalis, A.; Pappu, R. V., *J. Comput. Chem.* **2009,** *30* (5), 673-699.
120. Pande, V. S., *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100* (7), 3555-3556.
121. Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K., *Biophys. J.* **2008,** *94* (1), 182-192.
122. Kratky, O.; Porod, G., *Recl. Trav. Chim.* **1949,** *68* (12), 1106-1122.
123. Zhou, H. X., *Biochemistry* **2004,** *43* (8), 2141-2154.
124. Bertagna, A.; Toptygin, D.; Brand, L.; Barrick, D., *Biochem. Soc. Trans.* **2008,** *36*, 157-166.
125. Zhou, H. X., *J. Phys. Chem. B* **2001,** *105* (29), 6763-6766.
126. Ashbaugh, H. S.; Hatch, H. W., *J. Am. Chem. Soc.* **2008,** *130* (29), 9536-9542.
127. Abeln, S.; Frenkel, D., *PLoS Comput. Biol.* **2008,** *4* (12), 7.
128. Karplus, M.; McCammon, J. A., *Nat. Struct. Biol.* **2002,** *9* (9), 646-652.
129. Wang, G. L.; Dunbrack, R. L., *Bioinformatics* **2003,** *19* (12), 1589-1591.
130. Sweet, R. M.; Eisenberg, D., *J. Mol. Biol.* **1983,** *171* (4), 479-488.
131. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., *J. Comput. Chem.* **2004,** *25* (13), 1605-1612.