J|A|C|S ARTICLES

Published on Web 10/06/2010

Modeling Intrinsically Disordered Proteins with Bayesian Statistics

Charles K. Fisher,[†] Austin Huang,[‡,§] and Collin M. Stultz*[,†,‡]

Committee on Higher Degrees in Biophysics, Harvard University, and Harvard-MIT Division of Health Sciences and Technology, Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, United States

Received July 1, 2010; E-mail: cmstultz@mit.edu

Abstract: The characterization of intrinsically disordered proteins is challenging because accurate models of these systems require a description of both their thermally accessible conformers and the associated relative stabilities or weights. These structures and weights are typically chosen such that calculated ensemble averages agree with some set of prespecified experimental measurements; however, the large number of degrees of freedom in these systems typically leads to multiple conformational ensembles that are degenerate with respect to any given set of experimental observables. In this work we demonstrate that estimates of the relative stabilities of conformers within an ensemble are often incorrect when one does not account for the underlying uncertainty in the estimates themselves. Therefore, we present a method for modeling the conformational properties of disordered proteins that estimates the uncertainty in the weights of each conformer. The Bayesian weighting (BW) formalism incorporates information from both experimental data and theoretical predictions to calculate a probability density over all possible ways of weighting the conformers in the ensemble. This probability density is then used to estimate the values of the weights. A unique and powerful feature of the approach is that it provides a built-in error measure that allows one to assess the accuracy of the ensemble. We validate the approach using reference ensembles constructed from the five-residue peptide met-enkephalin and then apply the BW method to construct an ensemble of the K18 isoform of the tau protein. Using this ensemble, we indentify a specific pattern of long-range contacts in K18 that correlates with the known aggregation properties of the sequence.

Introduction

Constructing accurate models for disordered proteins is a challenging task. This is due, in part, to the realization that any reasonable model of the structure of a flexible protein must include a description of the thermally accessible states of the protein as well as the relative stability of each state. This information is quite difficult to obtain in practice because the set of ensembles that agree with any given set of experimental observations is typically highly degenerate; i.e., there are multiple ensembles that reproduce a given set of experimental observations within experimental error. Moreover, attempting to enumerate all of the degenerate solutions is computationally prohibitive for systems of even modest size, yet even if one could, it is not clear how to make inferences from a large set of possible solutions. This problem is particularly relevant for intrinsically disordered proteins (IDPs)—a class of polypeptides that cannot be adequately described by a unique native structure under physiologic conditions.[1] Much interest in understanding IDPs, such as tau protein, has been generated due to their proposed role in the development of neurodegenerative disorders such as Alzheimer's and Parkinson's diseases.[2−11]

Previous methods for mitigating the problem of degeneracy can be classified into two, not mutually exclusive, categories. First, some methods aim to find the simplest ensemble that reproduces a given set of experimental measurements. These

[†] Harvard University.
[‡] Massachusetts Institute of Technology.
[§] Current address: Center for Computational Molecular Biology, Division of Infectious Diseases, Brown University, Providence, RI 02906-2051, U.S.

(1) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(2) Barghorn, S.; Zheng-Fischhofer, Q.; Ackmann, M.; Biernat, J.; Bergen, M. v.; Mandelkow, E. M.; Mandelkow, E. *Biochemistry* **2000**, *39*, 11714–11721.
(3) von Bergen, M.; Friedhoff, P.; Biernat, J.; Heberle, J.; Mandelkow, E. M.; Mandelkow, E. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5129–5134.
(4) von Bergen, M.; Barghorn, S.; Li, L.; Marx, A.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E. *J. Biol. Chem.* **2001**, *276*, 48165–48174.
(5) Yao, T.-M.; Tomoo, K.; Ishida, T.; Hasegawa, H.; Sasaki, M.; Taniguchi, T. *J. Biochem.* **2003**, *134*, 91–99.
(6) Jeganathan, S.; Bergen, M. v.; Brutlach, H.; Steinhoff, H.; Mandelkow, E. *Biochemistry* **2006**, *45*, 2283–2293.
(7) Mukrasch, M. D.; Markwick, P.; Biernat, J.; von Bergen, M.; Bernardo, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2006**, *129*, 5235–5243.
(8) Huang, A.; Stultz, C. M. *Future Med. Chem.* **2009**, *1*, 467–482.
(9) Mylonas, E.; Hacher, A.; Bernardo, P.; Blackledge, M.; Mandelkow, E.; Svergun, D. I. *Biochemistry* **2008**, *47*, 10345–10353.
(10) Fischer, D.; Mukrasch, M. D.; Biernat, J.; Bibow, S.; Blackledge, M.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *Biochemistry* **2009**, *48*, 10047–10055.
(11) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *PLoS Biol.* **2009**, *7*, 399–414.

ensembles may be generated by finding the smallest number of structures necessary to reproduce the experimental data,[12,13] by weighting the structures in a conformational library in a way that maximizes the information entropy,[14] or by introducing restraints into a potential energy function that biases the resulting set of structures to have calculated averages that agree with experiment.[15,16] The second category consists of methods that enumerate several degenerate ensembles and then analyze them for similarity. In this case, a global measure of similarity between ensembles can be used to decide whether different solutions can be clustered or local measures of similarity can be used to indentify features that are common to all models.[17] All of these strategies have features that make them conceptually attractive, and a number of insights have been gained from their application. Ultimately, however, none of these methods directly address the underlying degeneracy of the problem.

To make the degeneracy problem explicit, suppose we have an intrinsically disordered protein under a prespecified set of experimental conditions (e.g., physiologic pH, pressure, temperature, etc.). One typically models such a protein by first sampling a relatively large set of conformations that represent possible accessible states of the system, $\{s_1, ..., s_n\}$. A model for the IDP is then built by either (1) selecting a smaller subset of structures that give calculated experimental observables that agree with experiment or (2) applying population weights to each of the n structures such that agreement between calculated observables and experiment is ensured.[13−19] In practice, the former approach is a special case of the latter since selecting a subset of structures is equivalent to setting the population weights of the excluded structures to zero. Consequently, we say that a structural ensemble is fully specified when both the set of structures $\{s_1, ..., s_n\}$ and the corresponding population weights, $\vec{w} = \{w_1, ..., w_n\}$ are known, where w_i is the weight of structure s_i and $\sum_{i=1}^{n} w_i = 1$.

For any given IDP there is some set of "true" weights, $\vec{w}^{\mathrm{T}} = \{w_1^{\mathrm{T}}, ..., w_n^{\mathrm{T}}\}$, that is a function of the relative free energies of each of the n structures. In principle, these probabilities could be calculated a priori once the potential energy surface is known. However, given the approximate nature of the energy functions that are used for the analysis of biomolecules, the exact calculation of relative free energies remains problematic.[20,21] Instead, as stated above, the relative probabilities of the different structures of an IDP are usually chosen to ensure that experimentally determined quantities agree with quantities calculated from the ensemble. For example, suppose $m_{\mathrm{exp},i}$ is the experimentally determined chemical shift of atom i. The best fit weights are those that minimize the error:

$$\xi_{M_i}(\vec{w}) = (m_{\mathrm{exp},i} - \sum_{j=1}^{n} w_j m_i(s_j))^2 = (m_{\mathrm{exp},i} - E_{\mathrm{CS}}[m_i|\vec{w}])^2 \quad (1)$$

where $m_i(s_j)$ is the predicted chemical shift of the ith atom in structure s_j, which is typically obtained from established algorithms such as SHIFTX,[22] and $E_{\mathrm{CS}}[m_i|\vec{w}]$ denotes the expected ensemble average of the chemical shift.

A major problem in determining an appropriate set of weights is that there are generally several different sets of weights, say, $\vec{w}_1, ..., \vec{w}_N$, with $\vec{w}_i \neq \vec{w}_j$, such that $\xi_{M_i}(\vec{w}_l)$ is less than some threshold that defines reasonable agreement with experiment for all l. In this case, we say that the problem is degenerate and it is not possible to distinguish between the different possible solutions without making additional assumptions.

In this paper, we present a method for analyzing the relative population weights. Our approach uses Bayesian statistics to determine a probability distribution for the population weight of each conformation in the ensemble. This probability distribution is called the posterior density and is based on both theoretical and experimental information. By recasting the problem in a statistical framework, we combat the degeneracy problem by calculating quantitative measures of uncertainty. We validate the Bayesian weighting (BW) approach using reference ensembles for the five-residue peptide met-enkephalin as a model system and then use BW to construct an ensemble of the K18 isoform of tau protein. Using this ensemble, we indentify a specific pattern of long-range contacts in K18 that correlates with the known aggregation properties of the sequence.

Theory

Overview. Rather than trying to identify a single "best fit" set of weights, a Bayesian approach specifies a probability distribution for the population weight of each structure in the ensemble. This allows one to quantify the uncertainty in the parameters of the ensemble so that inferences can be made using standard statistical methods. The posterior probability density for the weights given the observed experimental data is determined from the Bayes theorem:[23]

$$f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m}) = \frac{f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w})\, f_{\vec{W}}(\vec{w})}{\int \mathrm{d}\vec{w}\, f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w})\, f_{\vec{W}}(\vec{w})} \quad (2)$$

where $\vec{m} = \{m_1, ..., m_z\}$ denotes the vector of z experimental measurements.

The prior distribution, $f_{\vec{W}}(\vec{w})$, is chosen to represent a priori knowledge about the weights, $\vec{w}$. The likelihood function, $f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w})$, describes the probability of observing the experimental data, $\vec{m}$, for a given weight vector, $\vec{w}$. Below we discuss each of these terms in detail.

Prior Distribution. Let $\{s_1, ..., s_n\}$ denote a set of nonredundant structures. While this condition is not required to use the algorithm to obtain a point estimate for the weights, it is necessary to interpret the uncertainty measures that we introduce later. An estimate for the population weights could be obtained from the Boltzmann distribution:

(12) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(13) Zhang, Q.; Stelzer, A. C.; Fisher, C. K.; Al-Hashimi, H. M. *Nature* **2007**, *450*, 1263–1268.
(14) Choy, W.-Y.; Forman-Kay, J. D. *J. Mol. Biol.* **2001**, *308*, 1011–1032.
(15) Simone, A. D.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810–3811.
(16) Vendruscolo, M. *Curr. Opin. Struct. Biol.* **2007**, *17*, 15–20.
(17) Huang, A.; Stultz, C. M. *PLoS Comput. Biol.* **2008**, *4* (8), e1000155.
(18) Bernardo, P.; Bertoncini, C. W.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(19) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. *Biophys. J.* **2007**, *93*, 2300–2306.
(20) Cecchini, M.; Krivov, S. V.; Spichty, M.; Karplus, M. *J. Phys. Chem, B* **2009**, *113*, 9728–9740.
(21) Park, S.; Lau, A. Y.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 134102.
(22) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215–240.
(23) Bolstad, W. M. *Introduction to Bayesian Statistics*; John Wiley and Sons: Hoboken, NJ, 2007.

$$w_i^{\mathrm{P}} = \frac{\mathrm{e}^{-U(\mathrm{s}_i)/k_{\mathrm{B}}T}}{\sum_{j=1}^{n} \mathrm{e}^{-U(\mathrm{s}_i)/k_{\mathrm{B}}T}} \tag{3}$$

where the "P" stands for prior and $U(\mathrm{s}_i)$ is the energy of structure i. In principle, one could use other types of a priori information to construct $\vec{w}^{\mathrm{P}}$ as well.

The simplest prior distribution that is centered on $\vec{w}^{\mathrm{P}}$ and has a variance of k^{-1} is the Gaussian distribution. In practice, a simple Gaussian is not ideal because our domain of integration is the n-dimensional simplex, $S^n \equiv \{\vec{w} | \sum_{i=1}^{n} w_i = 1 \text{ and } w_i \geq 0\}$, rather than $\mathscr{R}^n$. Consequently, to define the prior distribution, we use an isomorphic coordinate transformation, $h: S^n \rightarrow \mathscr{R}^{n-1}$, which maps each point on S^n to an $(n-1)$-dimensional Euclidean space.[24−26] To simplify the notation, we denote the ith component of $h(\vec{w})$ by h_i. Each coordinate h_i, for $i = 1, ..., n-1$, is given by[24−26]

$$h_i = \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^{i} \ln(w_j) - i \ln(w_{i+1}) \right) \tag{4}$$

With this convention, we define the prior density for the population weights to be

$$f_W(\vec{w}|k) \propto \left(\prod_{i=1}^{n} w_i \right)^{-1} \exp\left[-\sum_{i=1}^{n-1} k \frac{(h_i - h_i^{\mathrm{P}})^2}{2} \right] \tag{5}$$

where $\vec{h}^{\mathrm{P}} = (h_1^{\mathrm{P}}, ..., h_{n-1}^{\mathrm{P}})$ is the point in $\mathscr{R}^{n-1}$ that corresponds to $\vec{w}^{\mathrm{P}}$ and $(\prod_{i=1}^{n} w_i)^{-1}$ is the Jacobian of the coordinate transformation. This simplicial normal distribution is the analogue of a Gaussian distribution for vectors of weights.[24−26]

Ideally, one would choose the variance to reflect the accuracy of $\vec{w}^{\mathrm{P}}$, but given the uncertainties in the accuracy of the underlying potential energy function, this approach is not practical. Therefore, we treat the variance as a random variable, with distribution $f_K(k)$, and average over all possible values to arrive at the prior distribution:

$$f_{\vec{W}}(\vec{w}) = \int_0^{\infty} f_{\vec{W}|K}(\vec{w}|k)\, f_K(k)\, \mathrm{d}k \tag{6}$$

In practice, we choose $f_K(k)$ to be a uniform distribution over an interval (k_L, ∞), where $k_L > 0$ can be made small (we use $k_L = 10^{-3}$) to ensure that $\vec{w}^{\mathrm{P}}$ does not strongly bias the posterior density.

Likelihood Function. Likelihood functions that describe the uncertainty for each type of experimental measurement must be defined, e.g., the RDC, chemical shift, radius of gyration estimate, etc. For each given type of measurement we also model the associated likelihood with a Gaussian density function. For example, the chemical shift likelihood function is defined as

$$f_{M_i|\vec{W}}^{\mathrm{CS}}(m_i|\vec{w}) = [2\pi(\varepsilon_{\mathrm{CS}}{}^2 + \alpha_{\mathrm{CS}}{}^2)]^{-1/2} \exp\left[-\frac{(m_i - E_{\mathrm{CS}}[m_i|\vec{w}])^2}{2(\varepsilon_{\mathrm{CS}}{}^2 + \alpha_{\mathrm{CS}}{}^2)} \right] \tag{7}$$

where $E_{\mathrm{CS}}[m_i|\vec{w}]$ is the value of the chemical shift calculated from the ensemble, $\varepsilon_{\mathrm{CS}}{}^2$ is the experimental error and $\alpha_{\mathrm{CS}}{}^2$ is the error in predicting the chemical shift. We use the program SHIFTX to predict chemical shifts and define α_{CS} as the rms error between predicted and observed chemical shifts in folded proteins reported by Neal et al.[22] In our model, each experimental shift measurement is independent so the joint likelihood is the product of the individual likelihood functions.

For some proteins, other types of experimental data, such as RDCs and information about the average radius of gyration, R_{G}, are available, and likelihood functions for these measurements are developed using a similar formalism (see the Methods), yielding separate probability distributions for each type of experiment, i.e., $f_{\vec{M}|\vec{W}}^{\mathrm{RDC}}(\vec{m}|\vec{w})$ and $f_{M|\vec{W}}^{R_{\mathrm{G}}}(m|\vec{w})$. In this setting the joint likelihood function for all of the measurements is the product of the RDC, chemical shift, and R_{G} likelihood functions:

$$f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w}) = f_{M|\vec{W}}^{R_{\mathrm{G}}}(m^{R_{\mathrm{G}}}|\vec{w})\, f_{\vec{M}|\vec{W}}^{\mathrm{RDC}}(\vec{m}^{\mathrm{RDC}}|\vec{w}) \prod_{j=1}^{N_{\mathrm{CS}}} f_{M_j|\vec{W}}^{\mathrm{CS}}(m_j^{\mathrm{CS}}|\vec{w}) \tag{8}$$

where N_{CS} is the number of chemical shift measurements.

Analysis of the Posterior Distribution. Once the prior distribution and the experimental likelihood have been specified, the posterior distribution is calculated using eq 2. The Bayesian estimate for the weight of the jth structure is given by

$$w_j^{\mathrm{B}} \equiv \langle w_j \rangle_{\mathrm{BW}} = \int \mathrm{d}\vec{w}\, w_j\, f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m}) \tag{9}$$

Similarly, $\vec{w}^{\mathrm{B}}$ denotes the vector of Bayesian estimates for all structures in the ensemble.

To assess the performance of the method, it is useful to introduce a metric that quantifies how different two vectors of weights are. The metric we use is based on the Jensen−Shannon divergence (JSD) between two weight vectors, $\vec{w}^{\mathrm{a}}$ and $\vec{w}^{\mathrm{b}}$:

$$\Omega^2(\vec{w}^{\mathrm{a}}, \vec{w}^{\mathrm{b}}) = S\left(\frac{\vec{w}^{\mathrm{a}} + \vec{w}^{\mathrm{b}}}{2} \right) - \frac{1}{2} S(\vec{w}^{\mathrm{a}}) - \frac{1}{2} S(\vec{w}^{\mathrm{b}}) \tag{10}$$

where $S(\vec{w}) = -\sum_{i=1}^{n} w_i \log_2(w_i)$ is the information entropy.[27,28] While $\Omega^2(\vec{w}^{\mathrm{a}},\vec{w}^{\mathrm{b}})$ is not a true metric (it does not satisfy the triangle inequality), $\Omega(\vec{w}^{\mathrm{a}},\vec{w}^{\mathrm{b}}) = [\Omega^2(\vec{w}^{\mathrm{a}},\vec{w}^{\mathrm{b}})]^{1/2}$ is a metric[29] and has the property that $0 \leq \Omega(\vec{w}^{\mathrm{a}},\vec{w}^{\mathrm{b}}) \leq 1$, and $\Omega(\vec{w}^{\mathrm{a}},\vec{w}^{\mathrm{b}}) = 0$ if and only if $\vec{w}^{\mathrm{a}} = \vec{w}^{\mathrm{b}}$.[27,29]

The Bayesian estimate for the weights is a point estimate that is derived from the posterior distribution, $f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m})$. However, the posterior distribution itself provides of wealth of information that can be used to quantify the uncertainty of this estimate. A useful measure to quantify the uncertainty in the population weights is the posterior expected divergence:

$$\sigma_{\vec{w}^{\mathrm{B}}} \equiv \langle \Omega^2(\vec{w}^{\mathrm{B}}, \vec{w}) \rangle_{\mathrm{BW}}{}^{1/2} = \left[\int \mathrm{d}\vec{w}\, \Omega^2(\vec{w}^{\mathrm{B}}, \vec{w})\, f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m}) \right]^{1/2} \tag{11}$$

This statistic falls within the range $0 \leq \sigma_{\vec{w}^{\mathrm{B}}} \leq 1$ and is equal to zero if there is no uncertainty in the population weights. The expected divergence plays a role for vectors of weights similar to that of the standard deviation in Euclidean space.

(24) Aitchison, J.; Egozcue, J. J. *Math. Geol.* **2005**, *37*, 829–850.
(25) Egozcue, J. J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barcelo-Vidal, C. *Math. Geol.* **2003**, *35*, 279–300.
(26) Mateu-Figueras, G.; Pawlowsky-Glahn, V. *Commun. Stat.—Theory Methods* **2007**, *36*, 1787–1802.
(27) Lin, J. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
(28) Shannon, C. *Bell Syst. Tech. J.* **1951**, *30*, 56–64.
(29) Endres, D. M.; Schindelin, J. E. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860.

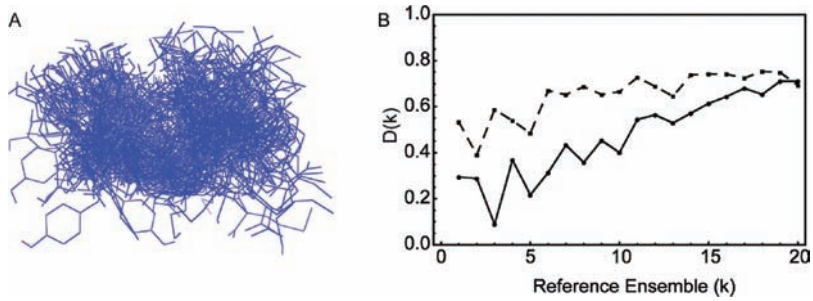

Figure 1. Degeneracy of point estimates for the reference ensembles: (A) diverse set of 95 structures for met-enkephalin constructed as described in the Methods; (B) average pairwise distances, $D_{NE}(k)$ (solid line) and $D_E(k)$ (dashed line), between the 10 solutions obtained with the optimization algorithm described in the Methods. The reference ensembles are ordered along the horizontal axis by increasing entropy.

Using this formalism, specific hypotheses can be tested quantitatively using Bayesian confidence intervals or model selection techniques.[23]

Results

Construction of Reference Ensembles. We tested the BW algorithm using the five-residue peptide met-enkephalin. Extensive replica exchange[30] simulations yielded 10 000 structures. To reduce this number to a more manageable size, a pruning algorithm was used to select low-energy structures that capture the structural diversity in the original set. This reduced set consists of 95 heterogeneous conformers (Figure 1A). Throughout this work we assume that this set of 95 structures is given and focus on the problem of weighting these conformations.

For each structure in this set, NMR chemical shifts were calculated for the Cβ, Cα, Hα, and backbone N−H and carbonyl atoms using the program SHIFTX,[22] yielding 28 chemical shifts per structure. Thus, the situation that we model in this paper is similar to the IDP case in that it is underdetermined; i.e., there are 94 degrees of freedom given by the weights (the condition on the sum of weights reduces the degrees of freedom by 1) and 28 experimental measurements.

Our goal is to determine whether the true conformational preferences in IDPs can be accurately inferred from a prior hypothesis for the population weights, $\vec{w}^{\mathrm{P}}$, and some set of experimental observables, $\vec{m} = \{m_1, ..., m_z\}$. To test this, we constructed a reference ensemble consisting of the set of 95 met-enkephalin structures and a prespecified set of "true" weights, $\vec{w}^{\mathrm{T}}$. The objective is to determine how well one can estimate this true set of weights given some experimental observations that have been made on the reference ensemble. The method of constructing reference ensembles as part of a validation strategy is well established in the literature, and useful insights have been obtained using this technique.[15,31]

To ensure that our results are not unduly influenced by the precise choice of $\vec{w}^{\mathrm{T}}$, we utilized 20 different sets of true weights, denoted as $\{\vec{w}^{\mathrm{T}_k}\}_{k=1}^{20}$. These weight vectors were chosen to guarantee that the various reference ensembles span a range of entropies. Since the entropy of a given weight vector quantifies the degree of structural heterogeneity in the ensemble, this ensures that the resulting reference ensembles span a range of structural disorder; i.e., high-entropy ensembles correspond to highly disordered states, while low-entropy ensembles have only a few conformations that have significant probability. Together the 95 structures and each true weight vector form a separate reference ensemble; hence, we have 20 different reference ensembles.

Degeneracy of Point Estimates. In this section our goal is to demonstrate that standard methods for finding optimal weights for an ensemble of structures yield degenerate solutions. These weights are typically found using non-Bayesian methods whose only goal is to optimize agreement with experiment; i.e., these methods are only concerned with optimizing eq 12 below and do not estimate the uncertainty in the underlying parameters of the model.

Traditionally, to model the conformational ensemble of an IDP, one searches for some weight vector, $\hat{w}$, that gives calculated average measurements (e.g., chemical shifts) that are similar to what is obtained from experiment; that is

$$\left[\frac{1}{z}\sum_{i=1}^{z}\xi_{M_i}(\hat{w})\right]^{1/2} \leq \varepsilon \tag{12}$$

where ξ_{M_i} is the error function, defined in eq 1, z is the number of experimental observations (e.g., number of chemical shifts), and ε is a reasonable estimate for the experimental error. We use $\varepsilon = 0.1$ for chemical shift measurements in proteins.[32,33] Simulated experimental NMR data for the kth reference ensemble, $\vec{m}^{\mathrm{T}_k} = (m_1^{\mathrm{T}_k}, ..., m_z^{\mathrm{T}_k})$, was created by calculating a set of measurements according to

$$m_i^{\mathrm{T}_k} = \sum_{j=1}^{n} m_{i,j}^{\mathrm{c}} w_j^{\mathrm{T}_k} + N(0, 0.1) \tag{13}$$

where $m_{i,j}^{\mathrm{c}}$ is the calculated chemical shift of residue i in structure j and $N(0,0.1)$ is a Gaussian noise term—having a mean of 0 and a standard deviation of 0.1 ppm—that is used to model typical experimental errors associated with chemical shift measurements in proteins.[32,33] This set of simulated experimental data was used to find weights that satisfy eq 12.

In addition to experimental error, one is often faced with the inability to calculate a given observable from a structure with perfect accuracy. This is the case, for example, with chemical shifts that are predicted using empirically derived algorithms.[22,34] To see how this uncertainty in predicting experimental measurements might affect the ability to reconstruct an ensemble from experimental data, we generated two sets of data.

(30) Okamoto, Y.; Fukugita, M.; Nakazawa, T.; Kawai, H. *Protein Eng.* **1991**, *4*, 639–647.
(31) Kuriyan, J.; Petsko, G. A.; Levy, R. M.; Karplus, M. *J. Mol. Biol.* **1986**, *190*, 227–254.
(32) Kurita, J.; Shimahara, H.; Utsunomiya-Tate, N.; Tate, S. *J. Magn. Reson.* **2003**, *163*, 163–173.
(33) Williamson, M. P.; Asakura, T. In *Protein NMR Techniques*; Reid, D. G., Ed.; Humana Press: Totowa, NJ, 1997; pp 53−69.
(34) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321–333.

To begin, we note that, in the world of our reference ensembles, the calculated chemical shift of the *i*th residue in the *j*th structure, $m^c_{i,j}$, corresponds to the result one would obtain if one could measure the corresponding chemical shift of that isolated conformation in solution. Algorithms that predict this chemical shift with 100% accuracy have no prediction error. We therefore refer to this case as the no error (NE) condition and define the predicted chemical shift in eq 1 to be $m_i(s_j) \equiv m^c_{i,j}$ and set $\alpha_{CS}{}^2 = 0$ in eq 7 (the rms error between predicted and observed chemical shifts). In the second case, we randomly perturbed the predicted chemical shifts using the reported SHIFTX error[22] by setting $m_i(s_j) \equiv m^c_{i,j} + \eta_i$, where $\eta_i \approx N(0,\alpha_i)$ in eq 1. In this case $\alpha_{CS}{}^2 \neq 0$ in eq 7 since this variable is determined by the published rms errors between SHIFTX predictions and the observed chemical shifts (e.g., for Cα carbons, $\alpha_{CS}{}^2 = 0.96$).[22] This scenario, which we refer to as the error-containing condition (E), models a more conservative view of the accuracy of the predicted chemical shifts. The simulated experimental data and the predicted chemical shifts of the structures were used with a simple non-Bayesian optimization algorithm described in the Methods to find weights that satisfy eq 12.

The non-Bayesian optimization algorithm was repeated 10 times for each reference ensemble, yielding 10 solutions for each reference ensemble in the no error (NE) condition and 10 solutions for the error-containing (E) condition. Hence, for each reference ensemble, the non-Bayesian optimization algorithm is repeated a total of 20 times. To assess the degeneracy of these solutions for each reference ensemble, we computed a degeneracy score that corresponds to the average pairwise distance from the 10 weight vectors for both the NE and E conditions. Given a set of solutions, $\{\vec{w}^i\}_{i=1}^{10}$, the average pairwise distance is given by $D_\lambda(k) = (\text{number of pairs})^{-1}\sum_{i<j}\Omega(\vec{w}^i,\vec{w}^j)$, where λ = NE or λ = E depending on what error condition was used to generate the set of solutions. We note that $D_\lambda(k)$ is 0 if and only if all of the solutions are identical.

As shown in Figure 1B, all of the reference ensembles have more than one unique solution; i.e., neither $D_{NE}(k)$ nor $D_E(k)$ is ever 0. Moreover, the high-entropy ensembles have the highest degeneracy scores, suggesting that all of the corresponding solutions are the most different. The situation is worse for E than for NE as $D_E(k) > D_{NE}(k)$ except for the highest entropy ensemble. This suggests that when the underlying ensemble is very inhomogeneous, accurate predictions for experimental observables do not help to limit the degeneracy of the problem. Moreover, since the results from separate runs of the optimization algorithm do not agree with each other, it is clear that simply finding a set of population weights that explains the experimental measurements is not sufficient to ensure the resulting ensemble is an accurate representation of the truth.

Validation of the BW Approach. In this section we will focus on the accuracy of $\vec{w}^B$ and the utility of $\sigma_{\vec{w}^B}$ as an estimate of the uncertainty in using $\vec{w}^B$ for an estimate of the true set of weights. The posterior distribution was calculated using eq 2 and then used to calculate the Bayesian estimate, $\vec{w}^B$, via eq 9, and the posterior expected divergence, $\sigma_{\vec{w}^B}$, via eq 11. Parts A and B of Figure 2 compare the accuracy, in terms of the JSD between the estimated weights and the weights of the reference ensemble (i.e., the true weights), of the BW method and an estimate obtained by numerical non-Bayesian optimization. Specifically, we compare $\Omega(\vec{w}^B,\vec{w}^T)$ to the minimum and maximum values of $\Omega(\vec{w}^O,\vec{w}^T)$ obtained from 10 independent runs of the optimization algorithm for each reference ensemble,

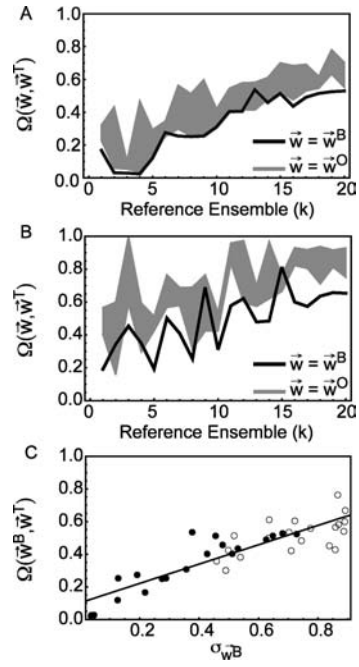

Figure 2. Validation of the BW method with reference ensembles. (A) and (B) compare the error in the Bayesian estimate, $\vec{w}^B$ (black line), to the error in the estimates obtained by non-Bayesian optimization, $\vec{w}^O$ (gray area), for the NE and E conditions, respectively. The *y* axis corresponds to the divergence between the solution and the true weight. The bottom and top of the gray area are determined by the minimum and maximum errors, respectively, of the 10 point estimates that represent solutions to eq 12. (C) The posterior expected divergence is correlated with the actual error between the Bayesian estimate and the true population weights. Solid circles are the NE results, and open circles are the E results. The best linear fit ($y = 0.1 + 0.6x$) with correlation $R = 0.88$ is shown as a solid black line.

where $\vec{w}^O$ is an estimate obtained from the non-Bayesian optimization. Our results suggest that the Bayesian point estimate is typically more accurate than point estimates obtained from an optimization algorithm that only ensures that the resulting solutions agree with experiment, i.e., that each solution satisfies eq 12.

Although the Bayesian estimate is generally more accurate than what one would obtain by optimizing eq 12 alone, we note that $\Omega(\vec{w}^B,\vec{w}^T)$ is generally not close to zero, especially for the high-entropy ensembles. This is expected when the posterior distribution has a large spread, in which case no point estimate will be able to adequately represent the distribution. The spread of the posterior distribution can be expressed using the expected divergence, $\sigma_{\vec{w}^B}$. As shown in Figure 2C, there is a strong correlation ($R = 0.88$) between $\sigma_{\vec{w}^B}$ and the divergence between the truth and the Bayesian estimate. This suggests that one can tell how accurate the Bayesian estimate is from $\sigma_{\vec{w}^B}$. Since $\sigma_{\vec{w}^B}$ is calculated directly from the BW algorithm, without knowledge of $\vec{w}^T$, our method provides a built-in error check on the population weights. In other words, the Bayesian estimate for the population weights is not always a good representation of the true ensemble, but we can specifically indentify these cases where the estimate significantly diverges from the truth. This is a unique feature of the BW approach; we do not simply obtain an estimate for the population weights but also an estimate of their uncertainty. Furthermore, we stress that the larger the value of $\sigma_{\vec{w}^B}$ the more important it is to summarize data with confidence intervals rather than point estimates. The ability to calculate interval estimates is another unique feature of the BW method.

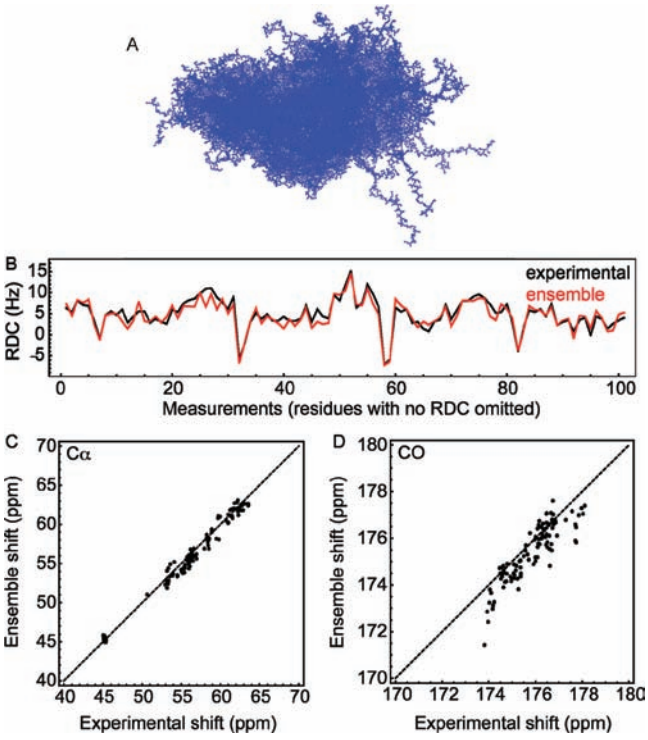

Figure 3. Application of the BW method to the K18 isoform of tau. (A) A diverse set of 300 structures was constructed as described in the Methods. (B) An overlay of the RDCs predicted from the ensemble and obtained from experiment shows good agreement ($R = 0.94$ and rms = 1.31 Hz). The predicted RDCs are uniformly scaled to account for uncertainty in predicting the magnitude of alignment. (C, D) The Cα ($R = 0.99$ and rms = 0.74 ppm) and CO ($R = 0.88$ and rms = 0.72 ppm) chemical shifts obtained from the ensemble show good agreement with experiment.

Residual Structure in the K18 Tau Isoform. We illustrate the utility of Bayesian confidence intervals by analyzing long-range contacts in the K18 isoform of tau protein. We used the BW algorithm to construct an ensemble of the 130-residue K18 isoform of tau protein using NMR chemical shifts, RDCs,[7,11,35] and the ensemble averaged radius of gyration determined by SAXS.[9]

We generated a set of energetically favorable structures for K18 by first dividing the protein into overlapping segments eight residues long. Extensive replica exchange simulations were performed to fully sample a wide range of structures for each segment. Structures for the full protein were then generated by joining the segments together, followed by energy minimization (see the Methods). (A similar procedure was previously used to explore the folding of peptide fragments in folded proteins.[36]) This yielded a set of 30 000 structures, which was then pruned to a set of 300 structures that again largely captured the structural heterogeneity in the original set (Figure 3A).

Application of the BW algorithm yielded an expected divergence of $\sigma_{\vec{w}^B} = 0.33$ corresponding to $\Omega^2(\vec{w}^B,\vec{w}^T) \approx 0.1$ bits based on the regression obtained with the reference ensembles (Figure 2C). This suggests that the posterior density is reasonably peaked. To provide some intuition for this number, a Jensen−Shannon divergence, Ω^2, of 0.1 corresponds to the difference between the weight vectors $\vec{w}^a = \{0,1\}$ and $\vec{w}^b = \{0.2,0.8\}$ in an ensemble consisting of just two structures.

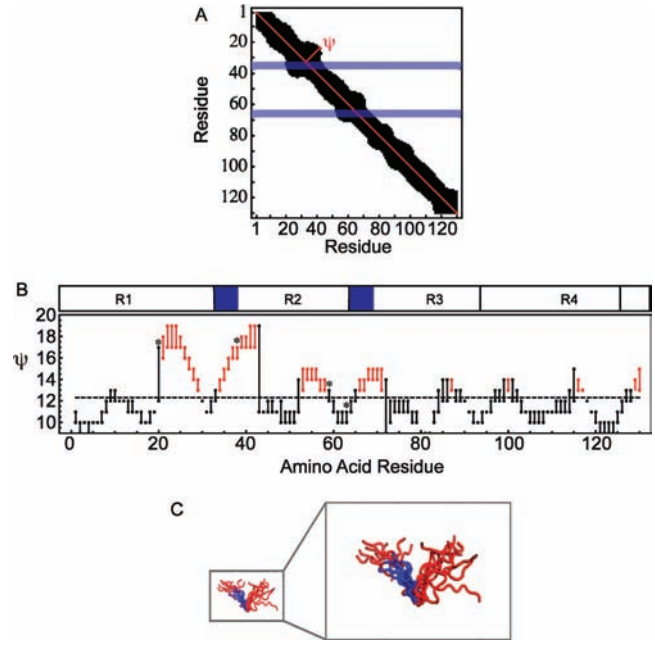

Figure 4. Analysis of long-range contacts in the K18 ensemble. (A) Contact map for K18 calculated from the Bayesian estimate for the weights. A black square indicates that the residues are within 25 Å on average. $\psi(i)$ is the length along the sequence to the farthest residue that makes a contact with residue i and is shown in red as the distance from the diagonal. PHF6* (residues 33−38) and PHF6 (residues 64−69) are highlighted in blue. (B) shows the 95% confidence intervals for ψ for each Cα in K18. The average ψ of all the residues is shown as a dashed line, and residues with confidence intervals that lie above this line are shown in red. The PHF initiating hexapeptides are shaded blue in the map of the sequence, and sites of mutation or phosphorylation known to alter the aggregation propensity of K18 in vitro are marked with an asterisk in the graph. (C) An overlay of the 10 most probable structures in the ensemble aligned via residues 20−44 (red), with PHF6* colored blue, illustrates a turn motif.

The resulting Bayesian estimate, $\vec{w}^B$, yields RDCs that are in very good agreement with experiment (Figure 3B). In addition, the average radius of gyration of the ensemble is about 36 ± 0.6 Å, compared to the experimental value of 38 ± 3 Å, and the agreement between the predicted and experimental chemical shifts is on the order of the SHIFTX[22] accuracy as shown in Figure 3C,D.

We analyzed the ensemble to look for long-range contacts in K18. A previous study analyzed long-range contacts in the 441-residue htau40 isoform using NMR paramagnetic relaxation enhancements (PREs).[11] Given that such experiments typically identify contacts up to 25 Å from the spin-label, we defined a contact as two residues that are within an average distance of 25 Å as this enables us to compare our data with those from previous experiments.[11]

Figure 4A shows a contact map constructed using the 300 structures in the K18 ensemble together with the Bayesian estimate of the weights, $\vec{w}^B$. Most of the inter-residue contacts occur between residues that are relatively close in the primary sequence. However, the regions near the paired helical filament (PHF) aggregation initiating hexapeptides PHF6* (residues 33−38) and PHF6 (residues 64−69) each make contacts with N-terminal residues that are relatively distant in the primary sequence. Interestingly, these regions are believed to be important for initiating tau aggregation in solution.[2−4]

While these data are interesting, we recognize that since $\sigma_{\vec{w}^B} \neq 0$, conclusions based only on an analysis of $\vec{w}^B$ may be misleading. Therefore, to account for the spread in the posterior

(35) Fischer, D.; Mukrasch, M. D.; Bergen, M. v.; Klos-Witkowska, A.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *Biochemistry* **2007**, *46*, 2574–2582.
(36) Ho, B. K.; Dill, K. A. *PLoS Comput. Biol.* **2006**, *2*, e27.

distribution, we constructed 95% confidence intervals for $\psi(i)$, a measure of how far along the sequence residue i makes contacts (Figure 4A). Figure 4B shows, in red, the residues that make long-range contacts using a 95% confidence interval. Interestingly, residues that are known to alter the aggregation potential of tau protein in vitro are located in regions that make relatively long range contacts. Furthermore, these data specifically highlight the two PHFs implicated in the tau aggregation process.[3] Looking at the 10 most probable structures in Figure 4C and zooming in on residues 20−40 shows that these contacts involve interactions between two extended regions separated by a turn formed by a PGGG sequence.

Discussion

The problem of degenerate conformational ensembles is difficult to overcome because the number of measurements that would be required to specify a unique ensemble typically pales in comparison to the number of measurements that are experimentally available. In this work, we demonstrated that the problem of degenerate conformational ensembles is particularly relevant for disordered proteins. In addition, we introduced an algorithm that allows one to manage degeneracy of the population weights within a coherent statistical framework. That is, for a given set of structures, prior weights, and experimental measurements, there is a unique posterior probability distribution on the space of population weights. An analysis of the posterior distribution using standard statistical techniques allows us to quantitatively summarize our knowledge about the structural ensemble.

Simulated experiments with met-enkephalin demonstrate that point estimates are often inadequate for making inferences about conformational preferences. This is especially true when there is error associated with calculating experimental observables from the structures; for example, it is clear from Figure 1B that for lower entropy ensembles improving the accuracy of algorithms for predicting chemical shifts would go a long way to reducing the degeneracy. In the case of higher entropy ensembles, such as those of IDPs, the degeneracy with accurate predictions for the experimental observables is already so large that having inaccurate predictions makes little difference.

The BW algorithm differs from previous methods in its ability to quantify uncertainty in the ensemble using $\sigma_{\vec{w}^B}$ and interval estimates. While the classical approach has only one criterion for a "good" ensemble, being agreement with the experimental data, we obtain a second criterion in terms of a small posterior expected divergence, $\sigma_{\vec{w}^B}$. That is, when $\sigma_{\vec{w}^B}$ is small, we can be confident that the ensemble is accurate, but if $\sigma_{\vec{w}^B} \gg 0$, more experimental data and more structures should be collected until the posterior expected divergence is minimized. Nevertheless, even in the case when $\sigma_{\vec{w}^B}$ is rather large, one can compute confidence intervals for the variables of interest that quantify the uncertainty in the relevant parameters.

After validating the BW algorithm using reference ensembles, we constructed an ensemble of the K18 isoform of tau protein. Tau is implicated in a number of neurodegenerative disorders, including Alzheimer's disease, through the formation of both soluble oligomeric states and insoluble aggregates known as neurofibrillary tangles.[2,4] K18 is the smallest isoform of tau, consisting of the four microtubule binding repeats that include two six-residue PHF initiating peptides—PHF6 and PHF6*—that are believed to be important for the aggregation process.[2−4] It is known that mutations at positions 38 (ΔK280), 59 (P301L), and 63 (S305N) result in dramatic increases in the aggregation propensity of both full-length tau and a variety of truncation mutants, including K18.[2−5] Furthermore, previous studies of K18 demonstrated that (pseudo)phosphorylation at position 20 (S262) leads to a conformational change that disrupts microtubule binding and decreases aggregation.[10,37] While position 38 is part of one of the PHF hexapeptides, positions 20, 59, and 63 are not; however, each of these residues occurs in one of the hot spots of long-range interactions or in the intervening turns. An analysis of the 10 most probable structures suggests that these turns are formed by PGGG sequences that preferentially occur toward the end of microtubule binding repeat regions (Figure 4C). Interestingly, it has been postulated that these PGGG motifs form turns at the end of regions that have a high propensity for the β-structure in the tau sequence.[38] Our data are in qualitative agreement with these findings and further suggest that the presence of these turns may play a role in modulating the aggregation propensity of tau.

Our findings suggest that mutation (or phosphorylation) of critical residues in K18 may alter the aggregation propensity of the peptide by affecting a network of long-range interactions. It has been postulated that phosphorylation at S20 decreases the aggregation propensity of tau by promoting electrostatic interactions with the end of R1 or beginning of R2, and our findings are in qualitative agreement with this hypothesis.[10] Moreover, our conclusions are in reasonable agreement with previous studies of the 441-residue htau40 isoform that found evidence of long-range contacts in the larger construct.[6,11] A recent FRET study found that the average distances between residues 49 (htau40 291) and 68 (htau40 310) (22 Å) and residues 68 (htau40 310) and 80 (htau40 322) (19 Å) in htau40 were less than the theoretical values for a random coil (about 36 Å).[6] We find that these average distances in K18 (30 and 31 Å, respectively) are also less than the theoretical random coil values, albeit a comparison of our data with the FRET data suggests that htau40 may be more compact than K18 in this region. In addition, an ensemble of htau40 constructed from simulations and PRE derived distances suggests the existence of long-range contacts between the end of R1 and the beginning of R2 as well as the end of R2 and beginning of R3 as we observe in K18.[11] The complementary results of these studies reinforce the notion that although tau is intrinsically disordered, it is not adequately described by a classic random coil.

In this work we focus on ensemble degeneracy with respect to the weights of a given set of structures. However, we recognize that there are two types of degeneracy that are associated with generating ensembles for intrinsically disordered proteins. First, there is the degeneracy in the weights of a given set of structures and then there is degeneracy with respect to the types of structures that are used to construct the ensemble. While this work deals with the former degeneracy problem, it is important to realize that the two types of degeneracy are not mutually exclusive problems. More precisely, the process of selecting a set of structures from a larger library to be part of the final ensemble is equivalent to assigning weights of zero to the unselected structures. In this sense the degeneracy problem with respect to the types of conformers that are included in an ensemble is a subset of the problem of assigning the correct weights to a larger ensemble.

(37) Schneider, A.; Biernat, J.; von Bergen, M.; Mandelkow, E.; Mandelkow, E. M. *Biochemistry* **1999**, *38*, 3549–3558.
(38) Mukrasch, M. D.; Biernat, J.; von Bergen, M.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *J. Biol. Chem.* **2005**, *280*, 24978–24986.

We further note that the BW method is not designed to outperform existing approaches in terms of agreement with experimental data or the ability to accurately reproduce reference ensembles. The unique value of the Bayesian approach lies in its ability to judge the accuracy of the constructed ensemble and in its ability to estimate the uncertainly in the model parameters and in macroscopic observables that are calculated from the model.

Prior to this study the accuracy of a given structural ensemble had been determined by assessing how well observables calculated from the ensemble agreed with their experimental counterparts. However, as our study clearly demonstrates, agreement with experiment alone does not guarantee that the associated ensemble is correct. Therefore, it is important to develop quantitative estimates of the uncertainty in the underlying model. In this regard, a Bayesian approach to estimating the relative stabilities of conformers in a structural ensemble has many attractive features. By providing quantitative estimates of the underlying uncertainty, the BW formalism provides a rigorous platform for generating confidence intervals for each of the parameters in the model. It is our view that such approaches provide a rigorous statistical framework for conducting hypothesis tests, and they help to assess what types of data and how much data are truly necessary to make confident inferences about the disordered protein of interest.

Methods

Construction of a Met-Enkephalin Structural Library. A 10 ns replica exchange molecular dynamics simulation was performed using the CHARMM force field and the EEF1 implicit solvent model.[39,40] Coordinates were saved every picosecond from the 300 K trajectory, resulting in a total structural library containing 10 000 structures. We then used a simple pruning algorithm to reduce the size of the structural library to a more manageable number. The algorithm consists of the following steps (iterated until convergence): (1) Pick two structures at random from the library. (2) If the root-mean-square deviation (rmsd) between the structures is less than a cutoff, then discard the structure with the higher energy. After pruning through the met-enkephalin structure library with an all-atom rmsd cutoff of 2.1 Å, we obtained a set of 95 representative structures.

Construction of a K18 Tau Structural Library. 1. Sampling Conformations of K18 Peptides. We generated a set of energetically favorable structures for K18 by first dividing the protein into overlapping segments eight residues long. A local sequence size of eight residues was chosen for the size of the peptides used in the segment simulations, which is approximately the size of the average persistence length of a polypeptide.[41] The sequence of K18 was divided into 26 peptides of 8 residues each, with an overlap of 3 residues between adjacent segments. A similar replica exchange protocol has been successfully used to sample conformations of eight residue peptides in a previous study.[36]

Each segment was simulated using 10 ns of replica exchange molecular dynamics using the EEF1 implicit solvent model.[40,42] The first 5 ns of REMD simulation was discarded as equilibration, and only the last 5 ns of simulation was used to draw conformations. Previous studies showed that the backbone entropy of peptides of this size typically equilibrates within 3.5 ns or less.[36] REMD simulations were run in heat baths exponentially spaced between 260 and 700 K. Exchanges were performed every 1 ps. Inspection of the REMD trajectories confirmed that exchanges frequently occurred between all temperatures. Structures are saved prior to each exchange, generating 5000 structures for each sequence segment sampled (a comparable number of structures are used in other stochastic models of the unfolded state).[41,43] Since 26 segments are required to cover the entire sequence of K18, 130 000 segment conformations are generated in total.

2. Constructing K18 Structures from Peptide Fragments. Structures of K18 were obtained by independently sampling and joining peptide conformations of local segments of the K18 sequence. This scheme is comparable to the structure-generation methods in statistical coil algorithms. However, instead of building sequence structures one residue at a time, the sequence is extended by independently sampling and adding one peptide segment at a time. Starting with the N-terminal segment, each subsequent segment structure is independently sampled from the REMD trajectory and aligned by the backbone atoms of the three overlapping residues. An individual K18 conformation is constructed as a PDB file is created with duplicate atoms erased and residues renumbered.

Structures were minimized to remove bad contacts using 1000 steps of steepest descent minimization followed by 1000 steps of adopted basis Newton−Raphson minimization. Inspection of the resulting structures showed that this minimization protocol removes bad contacts while preserving the overall topology of the K18 structure. We began evaluating the K18 structures by comparing the ensemble average radius of gyration to measured values obtained by SAXS.[44] Our set of structures model substantially underestimates the average radius of gyration of the ensemble, computing a radius of gyration of 1.81 nm, whereas the measured radius of gyration of K18 is 3.8 ± 0.3 nm. Therefore, we altered our protocol for generating K18 structures to ensure that they had an average radius of gyration that was similar to the experimental result. This was accomplished using an alternate procedure for selecting peptide fragments to be joined.

The new procedure favors selection of extended peptide conformations in the construction of K18 structures. Since we perform REMD simulations on each segment, we have 5000 structures for each segment, where the structures vary from the compact to the extended. A segment structure is chosen to be joined to the preceding segment according to the following probability distribution:

$$P(s_i) \propto e^{-\rho(Rg_i - Rg_E)^2} \tag{14}$$

where s_i is the ith structure from the REMD, $1 \le i \le 5000$, Rg_i is the backbone radius of gyration of peptide structure i, Rg_E is the backbone radius of gyration of a fully extended eight-residue peptide (8.5 Å), and ρ is the scaling parameter for favoring extended conformers. This formalism is equivalent to introducing a harmonic potential that is centered at the fully extended state with ρ as a force constant. For $\rho = 0$, this distribution reproduces the uniform sampling of conformers from the REMD simulation. By biasing the local conformational distributions toward more extended conformations, the distribution of the sampled K18 structures becomes more extended as well. A conformational library of 30 000 structures was constructed with 5000 structures each from $\rho \in \{0.00, 0.25, 0.50, 0.75, 0.875, 1.00\}$. A parameter value of $\rho = 0.875$ resulted in an ensemble with an average radius of gyration equal to the experimental measurement of 3.8 nm.

To reduce the size of the structural library to a number that could be easily run with the BW algorithm, the same pruning algorithm applied to met-enkephalin was used with K18, except we used a

(39) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
(40) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.
(41) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099.
(42) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(43) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(44) Mylonas, E.; Hascher, A.; Bernado, P.; Blackledge, M.; Mandelkow, E.; Svergun, D. I. *Biochemistry* **2008**, *47*, 10345–10353.

Cα-only rmsd cutoff of 18.2 Å. The rmsd cutoff was chosen to ensure the final set of conformations contained 300 structures, which was able to explain the experimental data and required a reasonable amount of computational resources.

BW Likelihood Function: Likelihood Function Definitions. We use a likelihood function for R_G similar to that for chemical shifts:

$$f^{R_G}_{M|W}(m|\vec{w}) = [2\pi\varepsilon_{R_G}{}^2]^{-1/2} \exp\left[-\frac{(m - E_{R_G}[m|\vec{w}])^2}{2\varepsilon_{R_G}{}^2}\right] \quad (15)$$

with the only difference being that R_G can be calculated exactly for each structure so there is no prediction error. Observables that are greater than zero, such as R_G, are usually modeled using a log-normal distribution. However, as long as the magnitude of the experimental error is much less than the magnitude of the actual measurement, a Gaussian distribution is a good approximation.

The RDC likelihood function in our model is

$$f^{\mathrm{RDC}}_{M_i|\vec{W},\Lambda}(m_i|\vec{w},\lambda) = (2\pi\varepsilon_{\mathrm{RDC}}{}^2)^{-1/2} \exp\left[-\frac{(m_i - \lambda E_{\mathrm{RDC}}[m_i|\vec{w}])^2}{2\varepsilon_{\mathrm{RDC}}{}^2}\right] \quad (16)$$

where $E_{\mathrm{RDC}}[m|\vec{w}]$ is the expected value of the RDC calculated from the ensemble, $\varepsilon_{\mathrm{RDC}}$ is the experimental error, and λ is a scaling factor to account for uncertainty in the magnitude of the predicted RDCs.[7] Because RDC prediction algorithms work by predicting the alignment tensor, and it is not clear how error in the orientation of the alignment tensor will propagate to the predicted RDCs, we have neglected uncertainty in the predicted RDCs for now. The joint likelihood function for N_{RDC} RDCs is

$$f^{\mathrm{RDC}}_{M|W}(\vec{m}|\vec{w}) = \int_{-\infty}^{\infty} f_\Lambda(\lambda) \prod_{i=1}^{N_{\mathrm{RDC}}} f^{\mathrm{RDC}}_{M_i|\vec{W},\Lambda}(m_i|\vec{w},\lambda)\, \mathrm{d}\lambda \quad (17)$$

where we choose $f_\Lambda(\lambda)$ to be a uniform distribution over an interval $(-\infty, \infty)$.

BW Monte Carlo Algorithm. A Markov chain Monte Carlo (MCMC) algorithm was used to calculate integrals of the general form of eq 9.[45−47] The posterior density given by eq 2 can be simulated using Gibbs sampling[48] by iteratively sampling a value of k, λ, and a set of weights from their conditional distributions and then discarding k and λ. The conditional distributions for k and λ can be sampled from exactly as they correspond to an exponential and Gaussian distribution, respectively. A Metropolis−Hastings step was implemented for sampling the weights using a simplicial normal distribution centered at the current weight vector as the proposal distribution. The proposal distribution had an isotropic variance that was tuned during an equilibration period so that about 25% of the steps were accepted.

To improve sampling of the posterior distribution a multiple-replica approach was employed. That is, several different Monte Carlo runs were performed in parallel on different processors. In the met-enkephalin simulations eight independent Markov chains (from the MCMC runs) were run at the same "temperature" ($T = 1$). For the Metropolis algorithm, adding a temperature parameter changes the acceptance probability from $\min(1, p(x')/p(x))$ to $\min(1, [p(x')/p(x)]^{1/T})$. The final sample was obtained by saving the weights from one of these chains selected at random in even intervals according to the prespecified sample size. This approach was modified to a replica exchange algorithm for the MCMC simulations for tau to improve mixing because of the larger number of structures.[49,50] The temperatures were exponentially spaced over the eight replicas between $T = 1$ and $T = 1.5$.[30,51] Swaps were attempted every 100 steps according to the "even−odd" exchange scheme with about 50% acceptance.[51,52] The weights from the low-temperature replica were saved in even intervals to match the prespecified sample size.

The met-enkephalin MCMC simulations consisted of a 5 million step mode search after which the system was restarted at the mode and equilibrated for another 5 million steps, followed by a sampling period of 50 million steps to yield a sample size of 20 000 weight vectors. The tau MCMC simulations consisted of a 100 million step equilibration period followed by a 1 billion step sampling period to yield a sample size of 50 000 weight vectors. The running averages for the Bayesian weight estimates and the posterior expected divergence were monitored to ensure that convergence was achieved. Experimental measurements consisted of Cβ, Cα, Hα, and backbone N−H and carbonyl chemical shifts,[35] backbone N−H RDCs,[7] and the radius of gyration.[9] Experimental errors were taken to be 0.1 ppm,[32,33] 1 Hz,[7,18] and 3 Å[9] for the chemical shifts, RDCs, and radius of gyration, respectively. Errors in the SHIFTX-predicted chemical shifts were taken from Neal et al.[22] The MCMC algorithm was implemented in C++ and is available from the authors upon request.

Non-Bayesian Optimization Algorithm. We used a simple evolutionary-based optimization algorithm to identify a set of weights for the 95 met-enkephalin structures that satisfy eq 12. This algorithm is based on a pairwise comparison selection mechanism that is commonly used in evolutionary game theory.[53] It searches the space of weights (i.e., the set of structures is fixed) through random mutation while the population "fitness" increases through natural selection. Each member of the population consists of a vector containing the weights of each of the 95 met-enkephalin structures. The algorithm began with 10 000 weight vectors (each vector contains 95 dimensions) drawn from a random distribution. At each step, two weight vectors, **A** and **B**, were selected at random from the population. A child vector, **C**, was drawn from a simplicial normal distribution centered about **A** with an isotropic variance of 0.1. Vector **C** replaced vector **B** if the error in **C** was less than or equal to the error in **B**, which corresponds to the low-temperature limit in the selection rule studied by Traulsen, Pacheco, and Nowak.[53] The process was repeated 1 million times, and the weight vector from this final set with the best agreement with the experimental data was saved. Thus, the final ensemble consisted of the 95 met-enkephalin structures and the best fit vector of weights.

Acknowledgment. We thank the Zweckstetter group for providing the NMR chemical shifts and RDC data for K18. This work was supported by NIH Grant 5R21NS063185-02.

JA105832G

(45) Chib, S.; Greenberg, E. *Am. Stat.* **1995**, *49*, 327–335.
(46) Hastings, W. K. *Biometrika* **1970**, *57*, 97–109.
(47) Metropolis, N.; Ulam, S. *J. Am. Stat. Soc.* **1949**, *44*, 335–341.
(48) Gelfand, A. E.; Smith, A. F. M. *J. Am. Stat. Soc.* **1990**, *85*, 398–409.
(49) Geyer, C. J. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*; Keramidas, Ed.; Interface Foundation: Fairfax Station, VA, 1991; pp 153−163.
(50) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(51) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2009**, *473*, 193–195.
(52) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.
(53) Traulsen, A.; Pacheco, J. M.; Nowak, M. A. *J. Theor. Biol.* **2008**, *246*, 522–529.