Comparative Studies of Disordered Proteins with Similar Sequences: Application to A#40 and A#42

Citation: Fisher, Charles K., Orly Ullman, and Collin M. Stultz. "Comparative Studies of Disordered Proteins with Similar Sequences: Application to Aß40 and Aß42." Biophysical Journal 104, no. 7 (April 2013): 1546–1555.

As Published: http://dx.doi.org/10.1016/j.bpj.2013.02.023

Publisher: Elsevier B.V.

Persistent URL: http://hdl.handle.net/1721.1/90574

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Comparative Studies of Disordered Proteins with Similar Sequences: Application to Aβ40 and Aβ42

Charles K. Fisher,[†] Orly Ullman,[‡] and Collin M. Stultz[†§*]

[†]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts; [‡]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts; and [§]Harvard-MIT Division of Health Sciences and Technology, Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics & Institute of Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT Quantitative comparisons of intrinsically disordered proteins (IDPs) with similar sequences, such as mutant forms of the same protein, may provide insights into IDP aggregation—a process that plays a role in several neurodegenerative disorders. Here we describe an approach for modeling IDPs with similar sequences that simplifies the comparison of the ensembles by utilizing a single library of structures. The relative population weights of the structures are estimated using a Bayesian formalism, which provides measures of uncertainty in the resulting ensembles. We applied this approach to the comparison of ensembles for Aβ40 and Aβ42. Bayesian hypothesis testing finds that although both Aβ species sample β-rich conformations in solution that may represent prefibrillar intermediates, the probability that Aβ42 samples these prefibrillar states is roughly an order of magnitude larger than the frequency in which Aβ40 samples such structures. Moreover, the structure of the soluble prefibrillar state in our ensembles is similar to the experimentally determined structure of Aβ that has been implicated as an intermediate in the aggregation pathway. Overall, our approach for comparative studies of IDPs with similar sequences provides a platform for future studies on the effect of mutations on the structure and function of disordered proteins.

INTRODUCTION

The defining characteristic of an intrinsically disordered protein (IDP) is that it populates a diverse set of conformations under physiological conditions (1,2). In principle, the volume of conformational space sampled by an IDP during its biological lifetime can be quite large. As a result, the number of independent experimental measurements that one can obtain for an IDP typically pales in comparison to the number of degrees of freedom that are associated with the disordered state. Nevertheless, it is possible to develop simplified models that capture the general features of an IDP ensemble by taking a coarse-grained view of conformational space. By coarse-graining we mean the process of dividing up the potentially infinite set of possible conformations into a finite number of discrete conformational states where each state represents a region of conformational space (3–9).

A number of important insights into the conformational properties of IDPs have been gained from methods designed to generate models of IDP ensembles that agree with experimental observations (3–9). Previous studies have shown, however, that the problem of generating an ensemble that agrees with experiment is frequently underdetermined even when the space of conformations has been coarse-grained (3,5,10). That is, one can typically find multiple different ensembles that reproduce the experimental data to within their associated uncertainties. Thus, developing methods that can quantify the uncertainty associated with a model of an IDP ensemble is an important task.

Methods from Bayesian inference provide a set of tools that can be used to model IDP ensembles, measure their uncertainties, and test hypotheses (3,5). To be precise, we define a coarse-grained ensemble as a finite set of structures, $S = \{s_1, \ldots, s_n\}$ (each representing a different conformational state), and their associated weights, $\vec{w} = \{w_1, \ldots, w_n\}$, where w_i is the population weight (or relative stability) of structure s_i. Assuming that the coarse-grained structural library S has been prespecified, the problem of modeling the ensemble reduces to estimating the vector of weights. In the Bayesian formalism, knowledge about the weights is expressed as a probability distribution that quantifies the uncertainty in the model of the ensemble. For example, if the experimental data are sparse, the variance of the probability distribution will be large. In general, our approach to constructing an ensemble—called Bayesian weighting (BW) or variational Bayesian weighting (VBW)—consists of four steps:

1. Generating a relatively large set of diverse but energetically favorable conformations;
2. Reducing the resolution of the set of structures through clustering;
3. Estimating the relative weights of the conformations in the structural library; and
4. Analyzing the ensemble (3,5).

Submitted August 28, 2012, and accepted for publication February 8, 2013.

*Correspondence: cmstultz@mit.edu

Editor: Martin Blackledge.

0006-3495/13/04/1546/10 $2.00

http://dx.doi.org/10.1016/j.bpj.2013.02.023

VBW is a computationally efficient approximation to BW; the procedure has the same steps and produces similar results (3,5). The diagram shown in Fig. 1 illustrates the main points of the approach.

In this work, we use several measured NMR observables including chemical shifts, residual dipolar couplings (RDCs), and $^3J_{HNH\alpha}$ scalar couplings to derive structural ensembles for two species of amyloid beta (Aβ)—Aβ40 and Aβ42. Although computational models for Aβ have been constructed and compared to experimental data previously, to our knowledge, this study is the first to directly use all of the available experimental data during model construction (11–15). To simplify comparison of the ensembles, we use a single library of conformations for both peptides so that any differences between the proteins were reflected solely in the population weights of the relevant structures. A Bayesian analysis of these data suggests that the relative probability of sampling soluble β-rich states, which may represent prefibrillar intermediates, is significantly greater for Aβ42 than Aβ40. Indeed, our approach gives a framework where we can make such statements with statistical confidence. Hence, this method offers a new, to our knowledge, paradigm for the comparison of unfolded ensembles of proteins that have similar amino-acid sequences and provides insights into the etiology of the difference in aggregation propensity between Aβ40 and Aβ42.

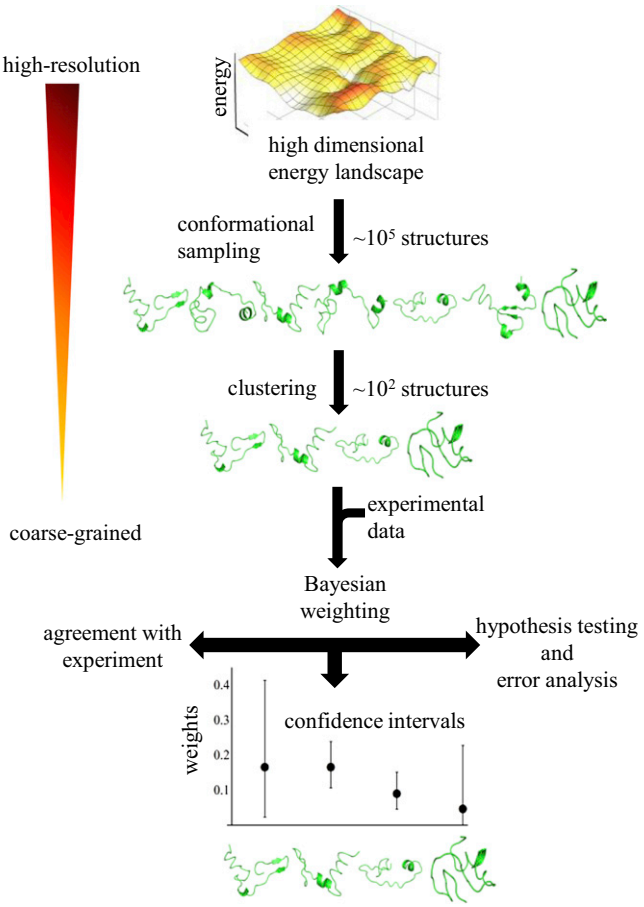

FIGURE 1 A schematic illustrating the construction of coarse-grained models for IDP ensembles using BW or VBW.

MATERIALS AND METHODS

Constructing the structural library

Conformational sampling was completed using a three-step process to achieve a heterogeneous set of structures. Half of the structures (i.e., 50,000) were obtained from a segment assembly procedure, and the other half of the structures (i.e., 50,000) were obtained from simulations of the full-length protein. The segment assembly process entailed breaking the Aβ42 sequence into overlapping segments and performing extensive conformational sampling of these segments using replica exchange molecular dynamics (REMD) simulations implemented in the software CHARMM (16) with the EEF1 (17) implicit solvent model. We used EEF1 here only because we have used it in prior work and have obtained fruitful results (4,18,19). Full-length Aβ42 conformations were generated by piecing the segments together one at a time, starting with the N-terminal segment, followed by energy minimization. This segment assembly protocol resulted in a total of 50,000 conformations.

The same REMD setup was used to sample conformations of full-length Aβ42, starting from a fully extended conformation of the protein. In addition, we performed quenched molecular dynamics on full-length Aβ42 to sample additional energetically favorable conformations. In total, 50,000 additional structures arose from simulations on the full-length protein. Full details of the segment assembly method, the REMD simulations, and the clustering method are in the Supporting Material.

Variational Bayesian weighting and calculating Bayes' factors

In variational Bayesian weighting (VBW), knowledge about the population weights of the conformations in the structural library is described using a Dirichlet distribution with a vector of parameters, $\vec{\alpha}$. The parameters of the Dirichlet distribution are found by minimizing Eq. 4. Once the parameters have been determined, the Bayes factor for comparing the weight of structure s_i in the Aβ40 and Aβ42 ensembles can be calculated as

$$BF(s_i) = \frac{P\left(w_i^{A\beta42} \geq w_i^{A\beta40}\right)}{1 - P\left(w_i^{A\beta42} \geq w_i^{A\beta40}\right)}. \quad (1)$$

The probability $P(w_i^{A\beta42} \geq w_i^{A\beta40})$ was estimated using the Laplace approximation (20,21) as described in the Supporting Material. According to Jeffreys' criterion, strong evidence for a significant difference in the weight of structure s_i between the Aβ40 and Aβ42 ensembles exists when $BF(s_i) \geq 10$ or $BF(s_i) \leq 1/10$ (22).

THEORY

Variational Bayesian weighting

In the following, we provide a brief overview of the VBW algorithm. More details are provided in Fisher et al. (3) and in the Supporting Material. The probability distribution for the relative weights of the conformations in the structural

library, conditioned on the observed experimental data, is determined using the Bayes theorem:

$$f_{\vec{W}|\vec{M},S}(\vec{w}|\vec{m},S) = \frac{f_{\vec{M}|\vec{W},S}(\vec{m}|\vec{w},S)f_{\vec{W}|S}(\vec{w}|S)}{f_{\vec{M}|S}(\vec{m}|S)}. \quad (2)$$

Here, $f_{\vec{M}|\vec{W},S}(\vec{m}|\vec{w},S)$ is the likelihood function, which describes the information obtained from the experimental measurements $\vec{m}$, and $f_{\vec{W}|S}(\vec{w}|S)$ is a prior probability distribution, which reflects a priori knowledge of the system. Because we did not have any a priori knowledge about the relative weights of the conformations in the structural library, we chose the noninformative Jeffreys' prior

$$f_{\vec{W}|S}(\vec{w}|S) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{w_i}}$$

(22,23). In this context, $f_{\vec{M}|S}(\vec{m}|S)$ is simply a normalizing constant (3).

The Bayesian approach obtains a probability distribution over all ways of weighting the structures in the structural library. This posterior density reflects the uncertainty in the weights that arises from having scarce or noisy experimental data, and from our inability to accurately calculate observables from a structure. Consequently, our analysis is specifically tailored to be extremely conservative in the treatment of uncertainties.

Calculating the moments of Eq. 2 typically requires extensive Monte Carlo simulations, which may take a long time to converge (3,18). Our approach to circumventing this problem is to use a simpler Dirichlet distribution with probability density function (24),

$$g(\vec{w}|\vec{\alpha}) = \frac{1}{\mathrm{B}(\vec{\alpha})} \prod_{i=1}^{n} w_i^{\alpha_i - 1}, \quad (3)$$

where the parameters, $\vec{\alpha}$, are chosen to minimize the distance between $g(\vec{w}|\vec{\alpha})$ and $f_{\vec{W}|\vec{M},S}(\vec{w}|\vec{m},S)$. Here, the normalizing constant $\mathrm{B}(\vec{\alpha})$ is the multinomial β-function. The Dirichlet distribution is an efficient choice for the approximating the posterior distribution shown in Eq. 2 because one can obtain an analytical expression for the KL divergence (a measure of distance) from $f_{\vec{W}|\vec{M},S}(\vec{w}|\vec{m},S)$ (3,25):

$$KL(\vec{\alpha}) = \int g(\vec{w}|\vec{\alpha}) \log\left(\frac{g(\vec{w}|\vec{\alpha})}{f_{\vec{W}|\vec{M},S}(\vec{w}|\vec{m},S)}\right) d\vec{w}. \quad (4)$$

Given that there is an analytical expression for Eq. 4, it is relatively simple to minimize the distance between the two distributions using a numerical optimization algorithm such as simulated annealing, as described in the Supporting Material. In addition, there are closed form expressions for the first two moments of the Dirichlet distribution, allowing for easy propagation of errors to characteristics calculated from the ensemble. For example, once we have identified the optimal set of parameters, $\{\widehat{\alpha}_i\}_{i=1}^{n}$, that minimizes Eq. 4, the Bayes estimates for the population weights are given by the simple formula

$$w_i^B = \frac{\widehat{\alpha}_i}{\sum_{j=1}^{n} \widehat{\alpha}_j}.$$

The best estimate of the ensemble corresponds to the structural library, $S = \{s_1, \ldots, s_n\}$, and the Bayes estimate of the population weights,

$$\vec{w}^B = \{w_1^B, \ldots, w_n^B\},$$

where

$$w_i^B = \int w_i g(\vec{w}|\vec{\alpha}) d\vec{w} = \frac{\widehat{\alpha}_i}{\sum_{j=1}^{n} \widehat{\alpha}_j}.$$

In general, this Bayes ensemble yields calculated observables that agree with the corresponding experimental data. However, as we have previously shown, agreement with experiment is insufficient to ensure that the ensemble is accurate (18). Consequently, we also use the Dirichlet distribution to calculate a normalized uncertainty parameter, $0 \le \sigma_{\vec{w}^B} \le 1$, that quantifies one's total uncertainty in the Bayes ensemble. In the limit $\sigma_{\vec{w}^B} \to 0$, the Dirichlet distribution approaches a Dirac delta function centered at the Bayesian weights. Our prior work suggests that in this scenario, the model is likely to be accurate; however, it is important to remember that the ensemble still only provides a coarse-grained representation of the conformational space of the protein (18). By contrast, when the uncertainty parameter is 1, it is likely that the ensemble is very inaccurate (18). However, even in the case when there is uncertainty, i.e., $\sigma_{\vec{w}^B} > 0$, we can calculate interval estimates for quantities of interest, i.e., confidence intervals. The ability to make quantitative statements about uncertainty in properties of the ensemble is one of the most important features of the BW and VBW algorithms. The ability to calculate confidence intervals also allows us to do rigorous hypothesis testing to find statistically significant differences between the ensembles.

RESULTS AND DISCUSSION

Model construction

Typically, if one is interested in comparing two ensembles, one begins by independently generating two structural libraries—i.e., one for Aβ40 ($S^{A\beta 40}$) and one for Aβ42 ($S^{A\beta 42}$). Next, each structural library could be separately

input into the VBW algorithm, along with the experimental data for the corresponding proteins, to obtain the associated Bayes estimates for the population weights, $\vec{w}^{B,A\beta40}$ and $\vec{w}^{B,A\beta42}$ (Fig. 2 *a*). Important differences between the two ensembles could be identified by comparing both the structures in the two ensembles along with their associated population weights (Fig. 2 *a*).

One challenge with this approach is that it is not straightforward to identify important differences between the ensembles if they were generated using different structural libraries (19,26). In this case, one has to choose a set of features that can be used for comparing the ensembles. For example, in a previous work, local structural motifs—regions consisting of six consecutive amino acids that adopt similar local structures in different global conformations—were used as features to compare K18 tau with the K18 ΔK280 mutant (19). In addition, Marsh and Forman-Kay (27) utilized global structural features such as the radius of gyration and secondary structure propensities to compare ensembles for the unfolded state of the drk N-terminal SH3 domain. Although there is merit in this approach, it is not always clear what features will yield useful insights.

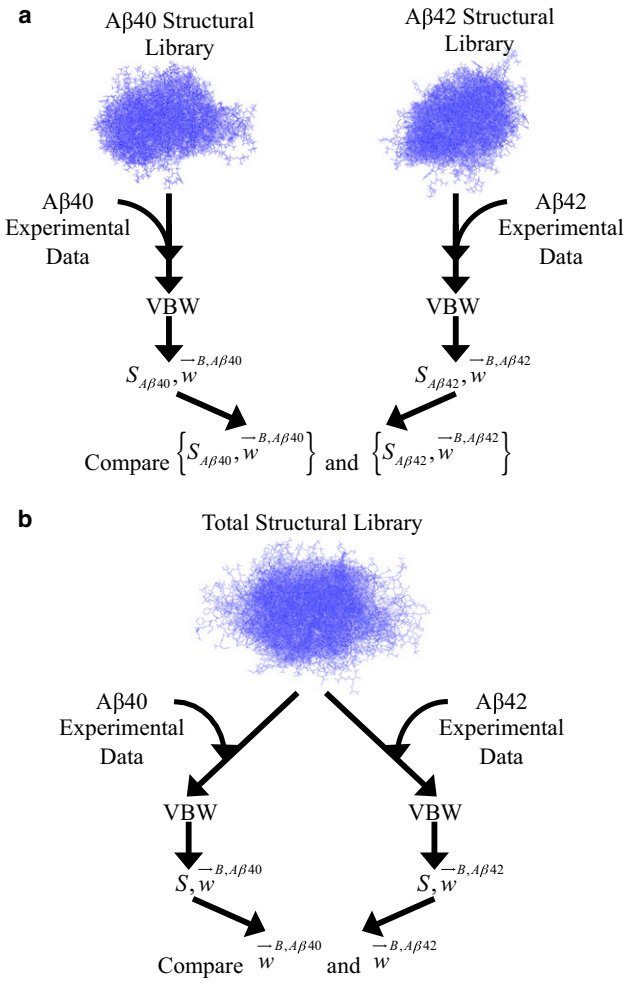

FIGURE 2 (*a*) Ensembles for Aβ40 and Aβ42 could be constructed independently, using different structural libraries, but comparing the resulting ensembles requires the difficult task of identifying important features, a priori. (*b*) Because the sequences of Aβ40 and Aβ42 are so similar, we assumed that a single structural library was adequate for describing the thermally accessible states for both peptides. With this assumption, the task of comparing the two ensembles is simplified to comparing the relative population weights of the structures. The ensemble shown at the top of panel *b* is a backbone alignment of all structures in the Aβ42 structural library.

By contrast, if we assume that, because their amino-acid sequences are so similar, Aβ40 and Aβ42 sample similar regions of conformational space then we can use the same coarse-grained structural library to construct ensembles for both proteins (Fig. 2 *b*). In this sense, we assume that the main difference between Aβ40 and Aβ42 is the frequency in which each protein samples the coarse-grained conformational states represented in the structural library. This provides an objective way to compare ensembles that does not require one to choose a set of important features. Instead, one can focus the analysis on conformations with large changes in stability.

Following this line of reasoning, we assumed that Aβ40 and Aβ42 sample similar conformational states and, therefore, constructed a single structural library for both Aβ42 and Aβ40 (after deleting the last two residues in the PDB file). Three different approaches were used to generate structures for the structural library:

1. A segment assembly method was used where the sequence of Aβ42 was divided into eight-residue-long overlapping segments resulting in eight segments covering the first 35 residues. (The last segment was seven residues long.) Adjacent segments overlapped by three residues. Each segment underwent REMD, and full-length Aβ42 conformations were generated by piecing together the segments arising from the REMD one at a time, starting with the N-terminal segment, followed by energy minimization.
2. Starting from a fully extended conformation of Aβ42, REMD simulations were run on the full-length protein.
3. Quenched molecular dynamics, again, were performed on the full-length protein.

Details of the construction of the structural library are provided in Materials and Methods and the Supporting Material. In brief, we generated 100,000 conformations using the methods mentioned above, followed by clustering based on overall secondary structure content, calculated with the software STRIDE (28). The final structural library consisted of 386 structures—each intended to represent a different conformational state. This number of conformations (i.e., 386) was determined by the number of unique combinations of total secondary structure content obtained from conformational sampling. Again, when applying the library to the Aβ40 sequence, the last two residues of every structure in the Aβ42 structural library were cleaved. An alignment of all 386 structures in the ensemble is shown

in Fig. 2 *b*. The structures within the ensemble represent a heterogeneous set of conformers that span a wide range of energetically favorable conformations and have variable secondary structure content and span a relatively wide range of radii of gyration (see Fig. S5 in the Supporting Material).

Experimental observables, specifically Cα, Cβ, and N chemical shifts, backbone NH RDCs and $^3J_{HNH\alpha}$ scalar couplings, were calculated for the conformations in the structural library (deleting the last two residues in the PDB file in the case of Aβ40) using the software SHIFTX (29) and PALES (30–32) and the Karplus equation with the parameters reported by Bruschweiler and Case (33), respectively. These predicted observables were used with their corresponding experimental measurements (12,15,34) to construct model ensembles for Aβ40 and Aβ42 using VBW.

Comparison with experimental data

A given experimental observable, E, corresponds to an ensemble average, $\langle E\rangle$. The ensemble average is given by

$$\langle E\rangle = \sum_{i=1}^{n} w_i f_E(s_i),$$

where $f_E(s_i)$ is the predicted experimental value arising from structure s_i. In light of this, any evaluation of an atomistic ensemble's ability to reproduce experimental data must account for both the experimental error as well as the inherent uncertainty associated with predicting experimental measurements from structural data. Consequently, the uncertainty in the ith experimental observable is the combined result of the experimental error, $\varepsilon_{i,\mathrm{exp}}$, and the error associated with the method used to predict the observable from a structure, $\varepsilon_{i,\mathrm{pre}}$. For example, the theoretical errors for chemical shift predictions were obtained from previously published assessments of the accuracy of chemical shift predictions (29,35), whereas the experimental errors were taken from Kurita et al. (36) and Williamson and Asakura (37). Of note, the errors associated with chemical shift predictions are roughly an order of magnitude larger than the associated experimental errors; e.g., the prediction error and experimental error for Cα chemical shifts are ~0.98 ppm and 0.1 ppm, respectively.

To be precise, the total error is formed from a combination of an experimental term ($\varepsilon_{i,\mathrm{exp}}$) and a theoretical term ($\varepsilon_{i,\mathrm{pre}}$), which reflects the intrinsic error in the algorithm used for predicting the ith measurement. In practice, the BW formalism uses the squares of the associated error to determine the variance of the likelihood function which is used to calculate the posterior density (see Eq. 2). Hence, the total, $\varepsilon_{i,T}$, error is defined by

$$\varepsilon_{i,T}^2 = \varepsilon_{i,\mathrm{exp}}^2 + \varepsilon_{i,\mathrm{pre}}^2. \quad (5)$$

Our use of Eq. 5 for representing the uncertainty in the observable quantities is a crucial part of our analysis because $\varepsilon_{i,\mathrm{exp}}^2 \ll \varepsilon_{i,\mathrm{pre}}^2$ for certain data types. For example, chemical shift errors typically dwarf the associated experimental errors, as noted above, such that $\varepsilon_{i,T}^2 \approx \varepsilon_{i,\mathrm{pre}}^2$. The total errors, $\varepsilon_{i,T}$, that we used for the Cα, Cβ, and N chemical shifts were 0.98, 1.16, and 2.43 ppm as described in Neal et al. (29) using SHIFTX. We note that errors associated with the prediction of RDCs from structure alone have not been systematically studied using a large dataset. Thus, we assumed that errors in the prediction of RDCs using PALES could be accounted for by uniformly scaling the predicted RDCs, and chose an experimental error of 0.45 Hz based on an RDC Q-factor of ~0.15 (38,39). We estimated the prediction error in the J-couplings using eight sets of Karplus constants reported in Bruschweiler and Case (33). The standard deviation in the predicted J-coupling, which is a function of the ϕ-angle, obtains a maximum of 0.85 Hz at $\phi = 60^\circ$. Assuming the same experimental error as the RDCs, we obtain a J-coupling error of

$$\sqrt{(0.45)^2 + (0.85)^2} \approx 0.95 \text{ Hz}.$$

This formalism, that accounts for the experimental and prediction error, is important because the relatively large uncertainty associated with prediction errors does not justify fitting to values that are much smaller than $\varepsilon_{i,T}$. Thus, the VBW objective function (see the Supporting Material) effectively weights deviations between the experimental data and the predicted data by $1/\varepsilon_{i,T}^2$ so that measurements with large uncertainties do not have much influence on the posterior distribution. In the end, the Aβ40 and Aβ42 Bayes ensembles obtained using VBW have root-mean-square errors from the experimental data that are on the order of $\varepsilon_{i,T}$ (Figs. 3 and 4). In addition, the average radius gyration of the both AB monomers is near 12 Å—a value that corresponds to ~16.8 Å using the relationship derived by Lindorff-Larsen et al. (40). Experimental measurements of the hydrodynamic radius of AB40 yield values of 16.8 Å (41).

Of all the experimental observables we consider, the errors associated with the $^3J_{HNH\alpha}$ scalar couplings are the poorest. Indeed, the correlation coefficients between the calculated and experimental data are poor (Figs. 3 and 4). Given that the prediction error for $^3J_{HNH\alpha}$ couplings (~1 Hz) is of the same order of magnitude as the actual values of the $^3J_{HNH\alpha}$ couplings (between 5.5 and 8 Hz), this is not surprising. We emphasize that given the inherent uncertainty in predicting J-couplings, ensuring that the error between the calculated values and the measured values is on the order of $\varepsilon_{i,\mathrm{exp}}$ alone is not warranted, and would lead to overfitting. Nevertheless, we note that the average error associated with $^3J_{HNH\alpha}$ couplings corresponds to an error in the associated ϕ-angle measurements of only 6° for Aβ40 and ~8° for

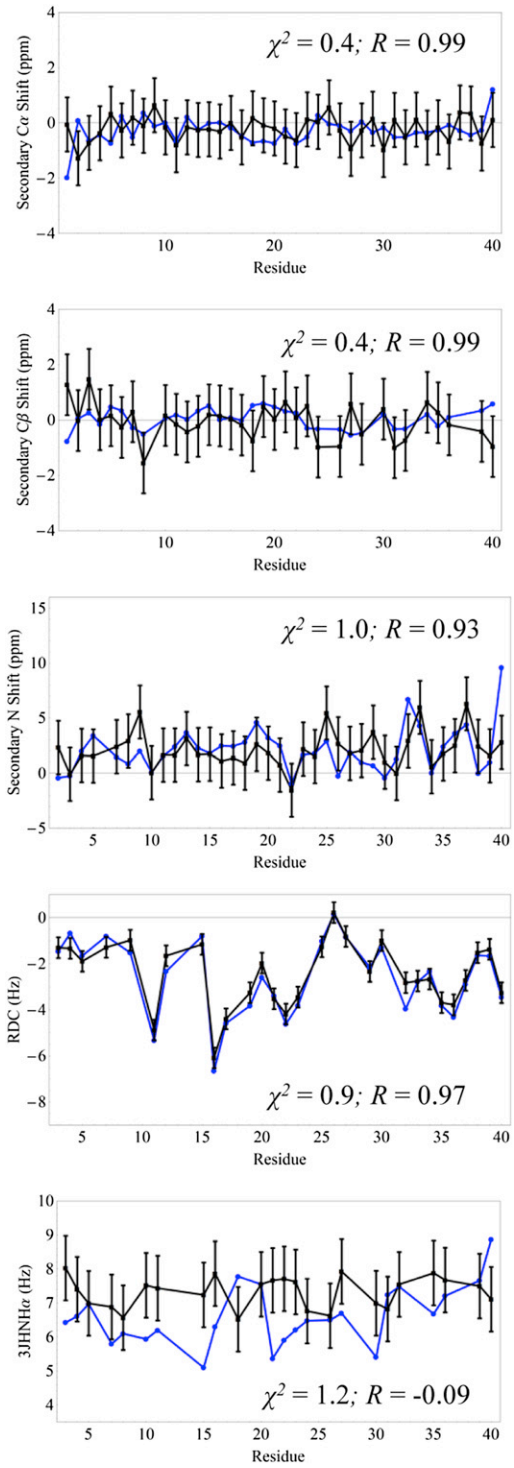

FIGURE 3 Agreement between experimental data (*blue*) and the data predicted from the Bayes ensemble (*black*) constructed for Aβ40. The error bars reflect a combination of experimental and prediction errors.

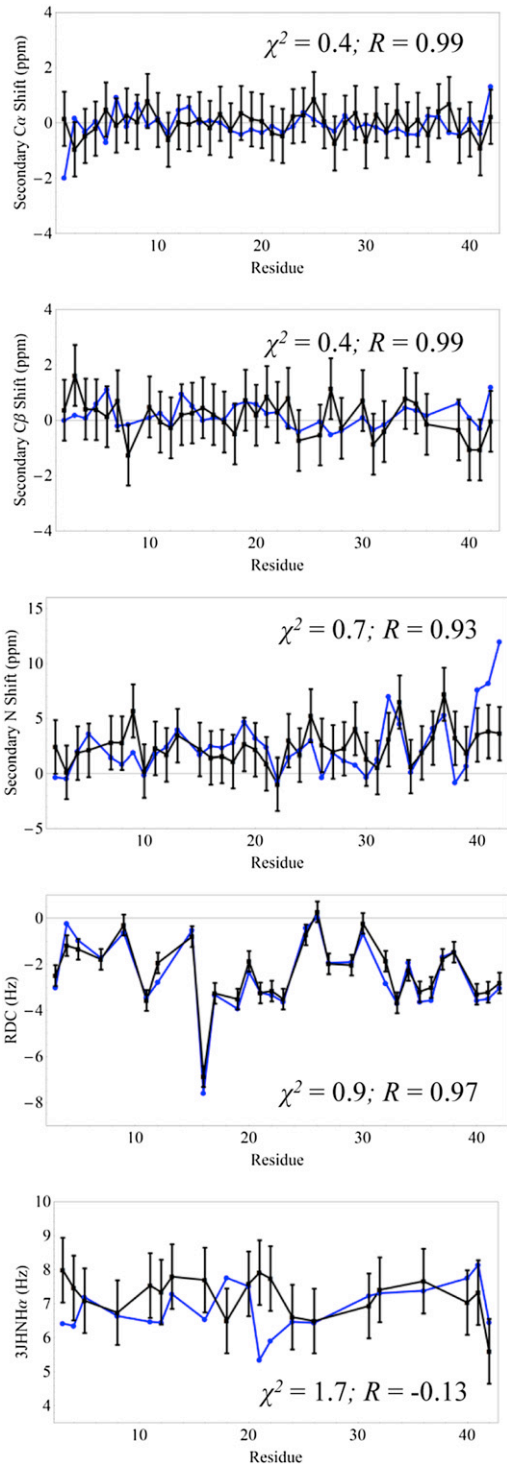

FIGURE 4 Agreement between experimental data (*blue*) and the data predicted from the Bayes ensemble (*black*) constructed for Aβ42. The error bars reflect a combination of experimental and prediction errors.

Aβ42 (Fig. 5); i.e., the $^3J_{HNH\alpha}$ scalar coupling errors correspond to relatively small absolute errors in the ϕ-angles. Consequently these data demonstrate that ensembles for Aβ40 and Aβ42 that agree with experiment can be generated from a single structural library.

Comparing ensembles for Aβ40 and Aβ42

The natural occurrence of multiple Aβ species with different aggregation propensities provides some clues about the mechanism of Aβ aggregation. It is well known that Aβ42

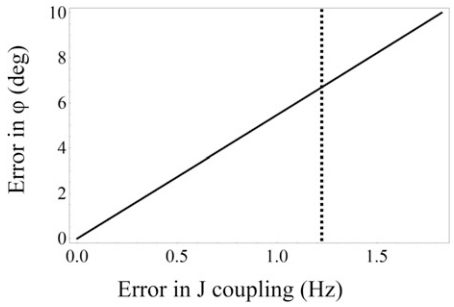

FIGURE 5 Relationship between errors in $^3J_{HNH\alpha}$ and the associated error in the ϕ-angle. Recall that the Karplus equation is $J(\phi) = A\cos^2(\phi - 60) + B\cos(\phi - 60) + c$. The error in the J-coupling, denoted Δ, as a function of the error in the ϕ-angle, denoted δ, was estimated using $\Delta(\delta) = \max_{\phi\in[-180,180]} |J(\phi+\delta) - J(\phi)|$, where $|J(\phi+\delta) - J(\phi)|$ is the given $^3J_{HNH\alpha}$ error. (*Dotted line*) Position associated with the average error between the calculated J-couplings and the measured values (and the corresponding error in the ϕ-angle) are shown.

has a higher propensity than Aβ40 for forming aggregates in vitro, is more toxic in vivo, and is more prevalent in senile plaques associated with Alzheimer's disease (42–44). It is not settled as to whether the toxic species corresponds to the fibrillar plaques or to smaller aggregates termed soluble oligomers; however, much of the recent research has concentrated on the soluble oligomeric species (45–52). Nevertheless, it is clear that Aβ self-association plays an important role in the pathogenesis of Alzheimer's disease. In this regard, a central question regarding Aβ is: how does such a small change in sequence—the amino-acid sequence of Aβ42 differs from that of Aβ40 only by the addition of an isoleucine and an alanine to the C-terminus—cause such a large change in aggregation propensity?

Structures rich in β-content have been proposed for both the fibrillar species of Aβ42 (53,54) and for soluble prefibrillar intermediates that may be involved in the formation of soluble oligomers (55). Nonetheless, a comprehensive understanding of the structural basis underlying the formation of Aβ aggregates requires knowledge of both the structure of the folded, aggregated state and the thermally accessible states of the unfolded protein. With regard to the unfolded monomeric state of the protein, a number of studies have used different methods, ranging from coarse-grained models to all-atom molecular dynamics simulations with explicit solvent, to model the conformational properties of Aβ40 and/or Aβ42 peptides (11–13,15,56–58). These studies have not identified any dramatic differences between the ensembles of Aβ40 and Aβ42. Instead, both Aβ peptides appear to adopt very heterogeneous ensembles that sample a variety of secondary structures, with a slightly higher propensity for β-sheet formation in Aβ42 than in Aβ40. Moreover, these observations have not provided quantitative estimates for the relative population of conformations containing β-structures that are similar to the aggregated conformers observed in prior studies (53,54).

The uncertainty parameters calculated from the models for the ensembles of Aβ40 and Aβ42 obtained from VBW were both ~0.6 (on a scale from 0 to 1), indicating a high degree of uncertainty in the population weights of the conformational states. Nevertheless, we can still calculate interval estimates for quantities that are calculated from the ensembles, which is one of the strengths of the BW formalism. Gross measures of the characteristics of the ensembles do not reveal any significant differences between Aβ40 and Aβ42; however, an analysis of the relative stabilities of the different coarse-grained structural states finds some important differences. The Bayes estimates for the weights differ for every conformation in the structural library, i.e., $w_i^{B,A\beta42} \neq w_i^{B,A\beta40}$ for every structure i, and the ratio of the structure population weights, $w_i^{B,A\beta42}/w_i^{B,A\beta40}$, ranges from 0.14 to 10.2. We performed Bayesian hypothesis tests to identify which of these differences were statistically significant. In total, out of the 386 different structures in the structural library, only four had strong evidence, according to Jefferys' criterion, i.e., Bayes' factors >10 or <0.1 (22), for statistically significant differences in their population weights between the Aβ40 and Aβ42 ensembles. We labeled the structures with significant differences s1–s4, with s1 and s2 having larger weights in the Aβ42 ensemble than in the Aβ40 ensemble, and s3 and s4 having larger weights in the Aβ40 ensemble than in the Aβ42 ensemble.

The three-dimensional structures of structures s1–s4, and the posterior probability distributions for their associated weights, are shown in Fig. 6. In each structure, residues 41 and 42 are colored yellow, to denote the fact that these residues are not present in Aβ40 structures. The width associated with each of the posterior distributions clearly illustrates that the weights of the structures cannot be uniquely determined. In other words, there is significant uncertainty about the weights, which is expressed by the standard deviations of these distributions. Nevertheless, we can use these data to make statistically rigorous statements about the relative population weights of each structure within the Bayesian formalism; i.e., there is strong evidence that structure s1 is more highly weighted in Aβ42 than it is in Aβ40 because the area in which the densities overlap is small (Fig. 6 *a*).

In sum, these data suggest that the ensembles representing monomeric Aβ42 and Aβ40 in solution contain conformations—like structure s1—that are relatively rich in β-structure. Indeed, according to the Bayes estimates for the weights, structure s1 is roughly 10 times more likely in the Aβ42 ensemble than in the Aβ40 ensemble. Moreover, this structure places the hydrophobic region involving residues 17–21 in position to interact with the hydrophobic C-terminus—which is consistent with the notion that the hairpin is stabilized by hydrophobic interactions that involve the C-terminal residues (34,59). Lastly, we note that an experimentally determined soluble β-hairpin was obtained using an affibody protein that selected for this conformer in Aβ40 (55). Moreover, it has been suggested that this structure

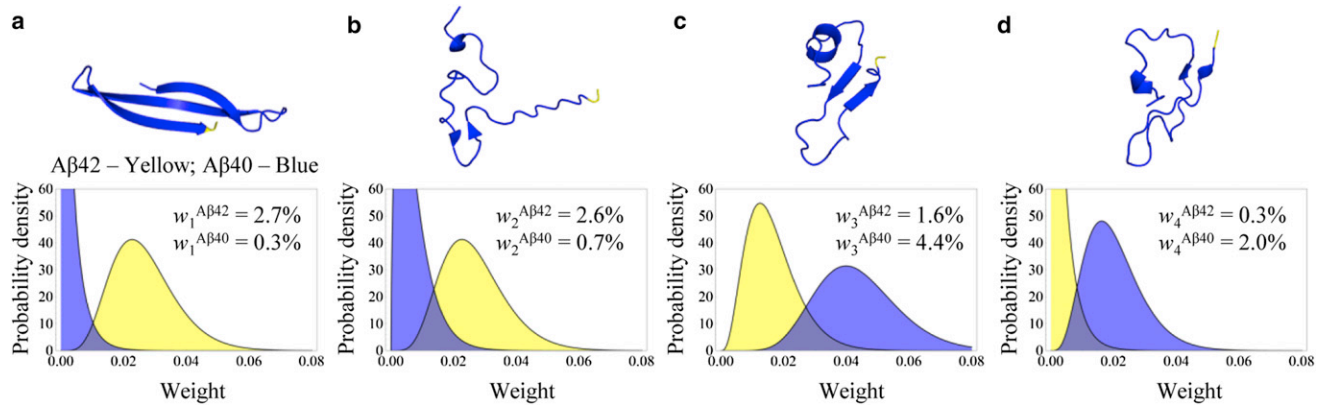

FIGURE 6 (*a*) Structure s1 (*top*), and the corresponding posterior probability distributions for its weight in the Aβ40 and Aβ42 ensembles (*bottom*). (*b*) Structure s2 (*top*), and the corresponding posterior probability distributions for its weight in the Aβ40 and Aβ42 ensembles (*bottom*). (*c*) Structure s3 (*top*), and the corresponding posterior probability distributions for its weight in the Aβ40 and Aβ42 ensembles (*bottom*). (*d*) Structure s4 (*top*), and the corresponding posterior probability distributions for its weight in the Aβ40 and Aβ42 ensembles (*bottom*).

represents a prefibrillar oligomeric hairpin intermediate (55). Interestingly, structure s1 is similar to the experimentally determined β-hairpin conformer, as shown in Fig. 7.

Finally, a recent article by Fawzi et al. (60) used novel dark-state exchange saturation transfer NMR experiments to directly probe the exchange between Aβ monomers and Aβ protofibrils. This study suggested a crucial role for the C-terminal residues in fibril formation, and highlighted differences in the dynamics of this region between protofibril bound Aβ40 and Aβ42. Overall, their results are consistent with the hypothesis that differences in the aggregation propensity of Aβ42 and Aβ40 can be linked to a higher population of hairpin conformations in Aβ42.

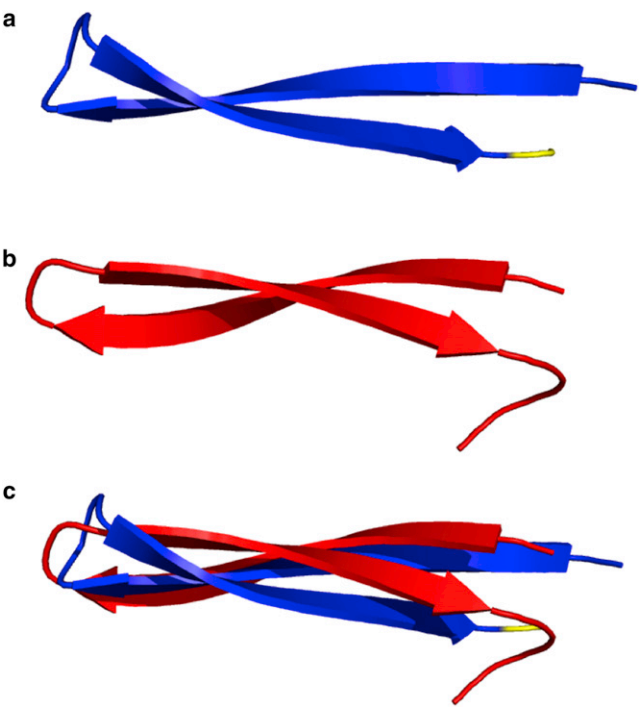

FIGURE 7 (*a*) Structure s1. (*b*) Structure of the experimentally determined β-hairpin conformer of Aβ (Model 1 of PDB:2OTK) (55). (*c*) Cα alignment of structure s1 and PDB:2OTK. The first 15 residues are not shown because these residues were disordered in the PDB:2OTK β-hairpin conformer.

CONCLUSIONS

In this work, we describe a Bayesian approach for generating coarse-grained models of IDP ensembles for comparative studies of proteins with similar sequences—in this case, the 40-and 42-residue species of Aβ. By ensuring that the ensembles shared the same library of conformations, structural differences between the peptides were reflected in the population weights of the relevant structures. The weights were estimated from experimental data using a Bayesian algorithm called variational Bayesian weighting (VBW) that accounts for the underdetermined nature of the problem by calculating a probability distribution over all ways of weighting the structures in the structural library. The standard deviations of the weights provide quantitative measures of how underdetermined the problem of weighting the conformations is, and allow one to propagate uncertainty to characteristics calculated from the ensemble using interval estimates.

Using Bayesian methods to estimate the population weights allowed us to perform hypothesis tests to identify conformations with statistically significant differences between their weights in Aβ40 and Aβ42. Bayesian hypothesis testing identifies four conformations with strong evidence indicating significant differences in their population weights in the two peptides; two of these conformations had higher weights in Aβ42 and two had higher weights in Aβ40. It is not surprising that only four out of 386 structures have statistically significant differences in their population weights between the two ensembles because the experimental data for Aβ40 and Aβ42 are similar and the associated errors (experimental + prediction) are large.

An analysis of these structures suggests that whereas both monomeric Aβ40 and Aβ42 sample soluble β-rich conformations that may be stabilized by hydrophobic interactions involving the C-terminal residues, Aβ42 appears to sample this conformation more readily. Moreover, the β-rich state in our ensemble is similar to the solution structure of an Aβ conformer that may play a role in both Aβ oligomerization and fibril formation (55). It has been suggested that fibril formation begins with the aggregation of β-hairpin-like structures into soluble oligomers, followed by a structural rearrangement that leads to cross-β structure (55). Indeed stabilization of this intramolecular β-hairpin may retard fibril formation in vitro (59,61). Nevertheless, we recognize that despite these observations, it is an open question whether the observed higher prevalence of the β-hairpin in Aβ42 completely explains the increased aggregation propensity of this protein. Indeed, it has been suggested that stabilization of the β-hairpin structure may form a therapeutic strategy for retarding the formation of Aβ fibrils (61).

One important consideration in these analyses is that the coarse-grained structural library must be prespecified, without recourse to the experimental data used in the Bayesian analysis of the population weights. Given that the equation for the posterior probability distribution is conditioned on the choice of structures, it is conceivable that a poor choice of structural library could lead to a poor quality model of the ensemble. To assess how a different choice of structures might impact our results, we performed additional analyses, described in the Supporting Material, of ensembles constructed for Aβ40 and Aβ42 using only the structures obtained from the segment assembly approach. These additional analyses support our primary conclusion that β-rich, prefibrillar conformations are more highly weighted in Aβ42 than in Aβ40 (see Fig. S3 and Fig. S4).

Thus, the models of Aβ obtained using VBW help to explain the difference in aggregation propensity of Aβ40 and Aβ42. Overall, our approach to constructing ensembles for comparative analyses can be used in future studies to examine the effects of mutations on IDP structure, function, and disease.

SUPPORTING MATERIAL

Methods, eight equations, and six figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(13)00240-3.

We thank Michael Zagorski and Chunyu Wang for providing the NMR data on Aβ40 and Aβ42 (12,15,34).

REFERENCES

1. Dunker, A. K., C. J. Oldfield, …, V. N. Uversky. 2008. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics.* 9(*Suppl 2*):S1.
2. Huang, A., and C. M. Stultz. 2009. Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease. *Future Med. Chem.* 1:467–482.
3. Fisher, C. K., O. Ullman, and C. M. Stultz. 2012. Efficient construction of disordered protein ensembles in a Bayesian framework with optimal selection of conformations. *Pac. Symp. Biocomput.* 2012:82–93.
4. Ullman, O., C. K. Fisher, and C. M. Stultz. 2011. Explaining the structural plasticity of α-synuclein. *J. Am. Chem. Soc.* 133:19536–19546.
5. Fisher, C. K., and C. M. Stultz. 2011. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 21:426–431.
6. Salmon, L., G. Nodet, …, M. Blackledge. 2010. NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132:8407–8418.
7. Mittag, T., J. Marsh, …, J. D. Forman-Kay. 2010. Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure.* 18:494–506.
8. Marsh, J. A., B. Dancheck, …, W. Peti. 2010. Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure.* 18:1094–1103.
9. Jensen, M. R., L. Salmon, …, M. Blackledge. 2010. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J. Am. Chem. Soc.* 132:1270–1272.
10. Ganguly, D., and J. Chen. 2009. Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states. *J. Mol. Biol.* 390:467–477.
11. Lin, Y.-S., G. R. Bowman, …, V. S. Pande. 2012. Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid β monomer. *Biophys. J.* 102:315–324.
12. Sgourakis, N. G., M. Merced-Serrano, …, A. E. Garcia. 2011. Atomic-level characterization of the ensemble of the Aβ(1–42) monomer in water using unbiased molecular dynamics simulations and spectral algorithms. *J. Mol. Biol.* 405:570–583.
13. Ball, K. A., A. H. Phillips, …, T. Head-Gordon. 2011. Homogeneous and heterogeneous tertiary structure ensembles of amyloid-β peptides. *Biochemistry.* 50:7612–7628.
14. Fawzi, N. L., A. H. Phillips, …, T. Head-Gordon. 2008. Structure and dynamics of the Aβ (21–30) peptide from the interplay of NMR experiments and molecular simulations. *J. Am. Chem. Soc.* 130:6145–6158.
15. Sgourakis, N. G., Y. Yan, …, A. E. Garcia. 2007. The Alzheimer's peptides Aβ40 and 42 adopt distinct conformations in water: a combined MD/NMR study. *J. Mol. Biol.* 368:1448–1457.
16. Brooks, B. R., R. E. Bruccoleri, …, M. Karplus. 1983. CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.
17. Lazaridis, T., and M. Karplus. 1999. Effective energy function for proteins in solution. *Proteins.* 35:133–152.
18. Fisher, C. K., A. Huang, and C. M. Stultz. 2010. Modeling intrinsically disordered proteins with Bayesian statistics. *J. Am. Chem. Soc.* 132:14919–14927.
19. Huang, A., and C. M. Stultz. 2008. The effect of a ΔK280 mutation on the unfolded state of a microtubule-binding repeat in tau. *PLOS Comput. Biol.* 4:1–12.
20. Shun, Z. M., and P. McCullagh. 1995. Laplace approximation of high-dimensional integrals. *J. R. Stat. Soc., B.* 57:749–760.
21. Strawderman, R. L. 2000. Higher-order asymptotic approximation: Laplace, saddlepoint, and related methods. *J. Am. Stat. Assoc.* 95:1358–1364.
22. Jeffreys, H. 1998. Theory of Probability. Clarendon Press, Oxford, UK.
23. Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London A Math. Phys. Sci.* 186: 453–461.
24. Rice, J. A. 2007. Mathematical Statistics and Data Analysis, 3rd Ed. Thomson Higher Education, Belmont, CA.

25. Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Stat.* 22:79–86.
26. Lindorff-Larsen, K., and J. Ferkinghoff-Borg. 2009. Similarity measures for protein ensembles. *PLoS ONE.* 4:e4203.
27. Marsh, J. A., and J. D. Forman-Kay. 2009. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J. Mol. Biol.* 391:359–374.
28. Heinig, M., and D. Frishman. 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* 32:W500–W502.
29. Neal, S., A. M. Nip, …, D. S. Wishart. 2003. Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *J. Biomol. NMR.* 26:215–240.
30. Zweckstetter, M. 2008. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* 3:679–690.
31. Zweckstetter, M., G. Hummer, and A. Bax. 2004. Prediction of charge-induced molecular alignment of biomolecules dissolved in dilute liquid-crystalline phases. *Biophys. J.* 86:3444–3460.
32. Zweckstetter, M. 2006. Prediction of charge-induced molecular alignment: residual dipolar couplings at pH 3 and alignment in surfactant liquid crystalline phases. *Eur. Biophys. J.* 35:170–180.
33. Bruschweiler, R., and D. A. Case. 1994. Adding harmonic motion to the Karplus relation for spin-spin coupling. *J. Am. Chem. Soc.* 116: 11199–11200.
34. Hou, L., H. Shao, …, M. G. Zagorski. 2004. Solution NMR studies of the Aβ (1–40) and Aβ (1–42) peptides establish that the Met35 oxidation state affects the mechanism of amyloid formation. *J. Am. Chem. Soc.* 126:1992–2005.
35. Kohlhoff, K. J., P. Robustelli, …, M. Vendruscolo. 2009. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* 131:13894–13895.
36. Kurita, J., H. Shimahara, …, S. Tate. 2003. Measurement of ^{15}N chemical shift anisotropy in a protein dissolved in a dilute liquid crystalline medium with the application of magic angle sample spinning. *J. Magn. Reson.* 163:163–173.
37. Williamson, M. P., and T. Asakura. 1997. Protein chemical shifts. *Methods Mol. Biol.* 60:53–69.
38. Zweckstetter, M., and A. Bax. 2000. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J. Am. Chem. Soc.* 122:3791–3792.
39. Berlin, K., D. P. O'Leary, and D. Fushman. 2009. Improvement and analysis of computational methods for prediction of residual dipolar couplings. *J. Magn. Reson.* 201:25–33.
40. Lindorff-Larsen, K., S. Kristjansdottir, …, M. Vendruscolo. 2004. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J. Am. Chem. Soc.* 126:3291–3299.
41. Danielsson, J., J. Jarvet, …, A. Graslund. 2002. Translational diffusion measured by PFG-NMR on full length and fragments of the Alzheimer Aβ (1–40) peptide. Determination of hydrodynamic radii of random coil peptides of varying length. *Magn. Reson. Chem.* 40:S89–S97.
42. Jarrett, J. T., E. P. Berger, and P. T. Lansbury, Jr. 1993. The C-terminus of the β-protein is critical in amyloidogenesis. *Ann. N. Y. Acad. Sci.* 695:144–148.
43. Selkoe, D. J. 1991. The molecular pathology of Alzheimer's disease. *Neuron.* 6:487–498.
44. Jarrett, J. T., E. P. Berger, and P. T. Lansbury, Jr. 1993. The carboxy terminus of the β-amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease. *Biochemistry.* 32:4693–4697.
45. Bernstein, S. L., N. F. Dupuis, …, M. T. Bowers. 2009. Amyloid-β protein oligomerization and the importance of tetramers and dodecamers in the etiology of Alzheimer's disease. *Nat. Chem.* 1:326–331.
46. Kuo, Y. M., M. R. Emmerling, …, A. E. Roher. 1996. Water-soluble Aβ (N-40, N-42) oligomers in normal and Alzheimer disease brains. *J. Biol. Chem.* 271:4077–4081.
47. Lambert, M. P., A. K. Barlow, …, W. L. Klein. 1998. Diffusible, nonfibrillar ligands derived from Aβ1–42 are potent central nervous system neurotoxins. *Proc. Natl. Acad. Sci. USA.* 95:6448–6453.
48. Watson, D., E. Castaño, …, A. E. Roher. 2005. Physicochemical characteristics of soluble oligomeric Aβ and their pathologic role in Alzheimer's disease. *Neurol. Res.* 27:869–881.
49. Ahmed, M., J. Davis, …, S. O. Smith. 2010. Structural conversion of neurotoxic amyloid-β (1–42) oligomers to fibrils. *Nat. Struct. Mol. Biol.* 17:561–567.
50. Sandberg, A., L. M. Luheshi, …, T. Härd. 2010. Stabilization of neurotoxic Alzheimer amyloid-β oligomers by protein engineering. *Proc. Natl. Acad. Sci. USA.* 107:15595–15600.
51. Yu, L., R. Edalji, …, E. T. Olejniczak. 2009. Structural characterization of a soluble amyloid β-peptide oligomer. *Biochemistry.* 48:1870–1877.
52. Ma, B., and R. Nussinov. 2010. Polymorphic C-terminal β-sheet interactions determine the formation of fibril or amyloid β-derived diffusible ligand-like globulomer for the Alzheimer Aβ42 dodecamer. *J. Biol. Chem.* 285:37102–37110.
53. Lührs, T., C. Ritter, …, R. Riek. 2005. 3D structure of Alzheimer's amyloid-β (1–42) fibrils. *Proc. Natl. Acad. Sci. USA.* 102:17342–17347.
54. Petkova, A. T., Y. Ishii, …, R. Tycko. 2002. A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci. USA.* 99:16742–16747.
55. Hoyer, W., C. Grönwall, …, T. Härd. 2008. Stabilization of a β-hairpin in monomeric Alzheimer's amyloid-β peptide inhibits amyloid formation. *Proc. Natl. Acad. Sci. USA.* 105:5099–5104.
56. Velez-Vega, C., and F. A. Escobedo. 2011. Characterizing the structural behavior of selected Aβ-42 monomers with different solubilities. *J. Phys. Chem. B.* 115:4900–4910.
57. Vitalis, A., and A. Caflisch. 2010. Micelle-like architecture of the monomer ensemble of Alzheimer's amyloid-β peptide in aqueous solution and its implications for Aβ aggregation. *J. Mol. Biol.* 403: 148–165.
58. Yang, M., and D. B. Teplow. 2008. Amyloid β-protein monomer folding: free-energy surfaces reveal alloform-specific differences. *J. Mol. Biol.* 384:450–464.
59. Mitternacht, S., I. Staneva, …, A. Irbäck. 2010. Comparing the folding free-energy landscapes of Aβ42 variants with different aggregation properties. *Proteins.* 78:2600–2608.
60. Fawzi, N. L., J. F. Ying, …, G. M. Clore. 2011. Atomic-resolution dynamics on the surface of amyloid-β protofibrils probed by solution NMR. *Nature.* 480:268–272.
61. Roychaudhuri, R., M. F. Yang, …, D. B. Teplow. 2012. Structural dynamics of the amyloid β-protein monomer folding nucleus. *Biochemistry.* 51:3957–3959.