Published in final edited form as:

Chem Phys. 2013 August 30; 422: . doi:10.1016/j.chemphys.2013.05.005.

Deducing conformational variability of intrinsically disordered proteins from infrared spectroscopy with Bayesian statistics

Anurag Sethi^{a,b}, Divina Anunciado^c, Jianhui Tian^a, Dung M. Vu^c, and S. Gnanakaran^{a,*}

^aTheoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA

^bCenter for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA

^cPhysical Chemistry and Applied Spectroscopy, Los Alamos National Laboratory, Los Alamos, NM, USA

Abstract

As it remains practically impossible to generate ergodic ensembles for large intrinsically disordered proteins (IDP) with molecular dynamics (MD) simulations, it becomes critical to compare spectroscopic characteristics of the theoretically generated ensembles to corresponding measurements. We develop a Bayesian framework to infer the ensemble properties of an IDP using a combination of conformations generated by MD simulations and its measured infrared spectrum. We performed 100 different MD simulations totaling more than 10 μ s to characterize the conformational ensemble of synuclein, a prototypical IDP, in water. These conformations are clustered based on solvent accessibility and helical content. We compute the amide-I band for these clusters and predict the thermodynamic weights of each cluster given the measured amide-I band. Bayesian analysis produces a reproducible and non-redundant set of thermodynamic weights for each cluster, which can then be used to calculate the ensemble properties. In a rigorous validation, these weights reproduce measured chemical shifts.

Keywords

Intrinsically disordered proteins; Fourier transform-infrared spectrum; Bayesian analysis; Conformational ensembles

1. Introduction

Intrinsically disordered proteins (IDPs) are a class of proteins that lack a stable native state in isolation under physiological conditions [1,2]. Instead, IDPs display a broad ensemble of conformations with varying amounts of residual secondary structure and solvent accessibility. IDPs are highly abundant in eukaryotic cells [3,4], and often function by folding upon binding to a target protein [5]. Some IDPs do not even fold in their functional state and serve as entropic chains in multivalent proteins [6,7]. In order to understand how IDPs function, it is essential to obtain atomistic level details of the ensemble of conformations they assume. Unfortunately, IDPs are very difficult to characterize, both

© 2013 Elsevier B.V. All rights reserved.

*Corresponding author. Address: MS K710 TA-3 T-6, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. Tel.: +1 505 665 1923; fax: +1 505 665 3493. gnanana@lanl.gov (S. Gnanakaran).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.chemphys.2013.05.005>.

experimentally and computationally, due to the heterogeneity in their conformational landscape. Of all the different experimental techniques that have been used to characterize disordered proteins, nuclear magnetic resonance (NMR) has provided the most comprehensive insights [8–10]. NMR has been used to monitor conformational changes in IDPs on a millisecond timescale [11]. It is desirable, however, to use infrared (IR) spectroscopy, as it allows characterization of secondary structure and conformational changes occurring at much faster timescales.

While IR spectroscopy has been applied to characterize and study the folding/unfolding of secondary structure elements in globular proteins [12,13], its application to IDPs has been limited. There have been a few studies that used Fourier transform IR spectrum (FT-IR), along with circular dichroism (CD), to characterize the amount of secondary structure present in IDPs [8]. The sensitivity of FT-IR to the backbone conformation of globular proteins is used to identify the peaks and widths in the spectrum of helices, strands, and random coils [14,15]. Specifically, the amide-I band in the 1600–1700 cm⁻¹ region is monitored for secondary structure content. This band mainly originates from the C=O vibrations of backbone carbonyl groups in peptide bonds and is sensitive to the environment of the carbonyl groups [16–19].

It is possible to use deconvolution of the amide-I band to infer the secondary structure content of large globular proteins, but this method is not straightforward. Typically, the amide-I band is decomposed into a sum of Gaussian or Lorentzian functions representing the different secondary structure elements in the protein [20]. However, the absorption spectra (and their peaks) for different secondary structural elements partially overlap with each other, and the specific peak position for each structural element can vary depending on its length and pattern of intra- and intermolecular hydrogen bonding [20]. Hence, a critical evaluation of the band assignment is necessary, and techniques such as Fourier deconvolution and second-derivative based methods have been developed to fit the spectrum to Gaussian functions [21–25]. Recently, singular value decomposition of the temperature-dependent IR spectrum was developed to determine the individual spectral components and their temperature dependence [26]. As IDPs exist in an ensemble of structures with varying residual secondary structure content and environment for the backbone groups, it is even more difficult to resolve the information in the amide-I band for large IDPs using these techniques.

Computational techniques have greatly enhanced the interpretation of the amide-I band and have provided underlying microscopic details leading to the observation of macroscopic amide-I band peaks and widths [27–31]. For example, molecular dynamics (MD) simulation presents an excellent technique to extract vibrational motions of biomolecules in the context of the surrounding aqueous environment, as it entails timescales and length scales relevant to amide-I bands [32]. Since it is impossible to perform accurate *ab initio* electronic structure calculations for a large number of conformations of peptides and proteins in an aqueous environment, the amide-I band is computed using a combination of classical MD simulations and empirical *ab initio*-based coupling and frequency “maps” [33]. The coupling of the amide-I mode from the different amide groups in the peptide causes splitting in fundamental frequencies of the chromophore (backbone amide group with predominant contribution from C=O stretching) [13]. As the coupling due to nearest neighbor occurs mainly through covalent bond interactions, these *ab initio* calculations have been used to calculate the dependence of nearest neighbor coupling terms on the backbone dihedral angles [30,34]. In contrast, the coupling due to all other chromophores occurs through space and is typically calculated using the transition dipole-transition dipole coupling model [13]. The frequency map takes into account the effects of the environmental perturbation, from solvent and ions, on the fundamental frequency of each chromophore. These shifts occur due

to the formation of intra-molecular and inter-molecular hydrogen bonds between the backbone atoms and their environment. Numerous maps were developed based on *ab initio* calculations on clusters of N-methylacetamide (NMA) in D₂O, which contains a single chromophore. In these maps, the frequency shift is assumed to be a linear combination of the solvent-induced electrostatic potential [35,36], electric field [37], or gradient of electric field [28] on the amide group atoms for each chromophore. These maps are assumed to be transferrable and have been used to study some proteins and membrane peptide bundles in an aqueous environment [27,29,31].

IDPs present a challenge to the above computational protocols, which seek to extract the underlying conformational ensembles that give rise to amide-I bands. In these protocols, nanosecond timescale MD simulations are performed on peptides or globular proteins to generate amide bands. [27]. In contrast, IDPs exhibit a large ensemble of conformations that may be stable for hundreds of nanoseconds or more [38,39]. The conformations can vary in their secondary structure content as well as in the extent of solvation of specific residues. Therefore, using a few long timescale conventional MD simulations to extract the structural characteristics of IDPs from the FT-IR spectrum can be difficult. The amide-I bands can only be obtained from MD simulations if the thermodynamic weights for the dominant states of IDPs can be estimated. Alternatively, amide-I bands can be analyzed using thermodynamically weighted ensembles of conformations using enhanced sampling methods such as replica exchange MD simulations [29]. However, such enhanced sampling methods can only be performed in practice for smaller proteins. For larger IDPs, convergence cannot be achieved with a few long timescale MD simulations using current computational resources [40]. The probability of exploring the relevant phase space of such an IDP increases if we use many shorter timescale simulations [39]. While the phase space of an IDP is efficiently explored using many short MD simulations, the thermodynamic information of the relative stability of each conformational state is lost using such an approach. Thus, a quantitative comparison of an amide-I band of a larger IDP requires two evaluations: (i) use numerous short MD simulations to sample the phase space effectively and capture different conformational states, and (ii) assign thermodynamic weights to these conformational states and calculate the amide-I band for each state.

In this study, we have developed a Bayesian framework to estimate the thermodynamic weights of an ensemble of conformations from the amide-I band of an IDP. This approach aids in estimating the mean and divergence (or redundancy) in the ensemble properties of an IDP that best fits the experimental amide-I band from a large number of conformational states generated from several hundred short MD simulations. Recently, a similar Bayesian formulation was developed to characterize the conformational properties of IDPs using NMR spectroscopy [41,42]. We apply our methodology to experimental and theoretical characterization of the linear IR spectrum of human α -synuclein, a 140-residue prototypical IDP, in water. Despite being highly studied, the conformational variability of α -synuclein in water, and its precise biological function, remain unknown. α -Stranded aggregates of α -synuclein are associated with Parkinson's disease [43,44], which is diagnosed postmortem by the presence of Lewy bodies in neurons. Several biophysical studies including NMR [45–50], electron paramagnetic resonance [51–53], fluorescence [54], and circular dichroism (CD) [55–58] spectroscopies have provided insights into some of the ensemble properties of α -synuclein in water. In addition, single molecule atomic force microscopy and single molecule fluorescence resonance energy transfer studies indicate that α -synuclein exists in an ensemble of different conformations [59,60]. The binding of phospholipids or detergents like sodium dodecyl sulfate (SDS) to α -synuclein induces it to fold into a helix-turn-helix or elongated helix conformation [61]. MD simulations indicate that α -synuclein rapidly collapses in an aqueous environment both in the absence and presence of SDS [39,62]. These studies indicate that the protein exists in an ensemble of collapsed and expanded

states, and the heterogeneous collapsed states lead to the formation of highly modular communities of contiguous peptides that can be studied independently of one another [39].

We describe the Bayesian formalism [41,42] for interpreting the IR spectrum (amide-I band) of an IDP in the methods section. We report and compare the measured and simulated amide-I band of α -synuclein in an aqueous environment. Using more than 100 MD simulations of 100 ns each for α -synuclein in explicit water, we illustrate the challenges faced while simulating amide bands from MD simulation trajectories due to the inherent conformational variability of an IDP. Then, we show how one could infer thermodynamic weights of different conformational states using the proposed Bayesian framework. This approach makes a quantitative comparison between the calculated amide-I band and the measured spectrum, and provides microscopic insights into the conformational variability of an IDP. In addition, we evaluate whether completely different ensembles of conformations of α -synuclein would also lead to a nearly identical amide-I band. Finally, we rigorously validate the predicted ensemble properties of α -synuclein by calculating NMR chemical shifts and compare them with measurements.

2. Methods

2.1. Experimental IR spectrum

2.1.1. Sample preparation—The pET-30a plasmid containing cloned wild type α -synuclein was expressed in *Escherichia coli* BL21 (DE3). An overnight culture of transformed cells was diluted 100-fold into LB media with kanamycin and grown at 37 °C. At OD A₆₀₀ of 0.5–0.6, the culture was induced with 1 mM IPTG and allowed to grow for 5–6 h. The cells from the culture were pelleted and resuspended in osmotic shock buffer (30 mM Tris–HCl, 40% sucrose, 2 mM EDTA-disodium, pH 7.2). After incubation at room temperature for 10 min, the sample was centrifuged. The pellet was resuspended in ice cold water with 4.0 M MgCl₂ and kept on ice for 3 min. The periplasmic proteins in the supernatant were collected by centrifugation. The sample was then dialyzed overnight against 30 mM Tris–HCl, pH 8.0. After centrifugation, the supernatant was collected and loaded onto a Q-Sepharose column. α -synuclein was eluted with a 0–500 mM NaCl gradient in 30 mM Tris–HCl buffer. The eluted fractions were analyzed by SDS–PAGE. The fractions that contained the protein were pooled and dialyzed against phosphate buffer. Further purification was performed using SEC column-FPLC. The collected fractions that contained the protein were analyzed by SDS–PAGE and pooled. The pooled protein sample was then dialyzed against water and lyophilized. The protein was analyzed using electrospray ionization mass spectrometry.

2.1.2. IR experiment—The purified protein was repeatedly lyophilized with D₂O (Cambridge Isotope Laboratories, Andover, Massachusetts) to allow for complete exchange of amide protons. The lyophilized protein was resuspended into a final buffer of 20 mM sodium phosphate (Na₃PO₄), and 100 mM NaCl at pH* 7.5 in D₂O. Na₃PO₄ was primarily used to maintain the pH of the solution. The pH* is the uncorrected (for D₂O) pH-meter reading at 25 °C. Sample concentrations of the protein solution used in the IR experiment were about 200 μ M based on an extinction coefficient of 5120 M^{−1} cm^{−1} at 280 nm. The buffer and the protein solutions were filtered to remove aggregates and used without further purification. The Fourier-transform infrared (FT-IR) spectra were obtained using a Thermo Nicolet Nexus 870 FT-IR spectrometer (Thermo Scientific, Waltham, Massachusetts). For the equilibrium measurement, a thermostatted, dual-compartment sample cell with CaF₂ windows and a 100 μ m Teflon spacer, was used to allow separate measurements of sample and reference spectra under identical conditions. The absorbance was computed using a single beam D₂O reference background obtained at the same temperature. We used the

standard spectral deconvolution Peak Resolve function from the Omnic 8.0 software (Thermo Scientific, Waltham, Massachusetts) to perform the peak fitting analysis of the IR spectrum. We note that while the protein concentration (200 μM) in which we performed these measurements may be conducive to aggregation or amyloid fibril formation, we believe that this protein is still mainly monomeric and soluble under these conditions.

2.2. MD simulations

We performed more than 100 all-atom MD simulations of α -synuclein in explicit water, with each simulation starting from a different initial condition. Each of these simulations ran for 100 ns. Details of these simulations were published previously [39]. Briefly, the initial conformations for half of those simulations were unfolded random coils generated using the random coil generator [63]. The initial conformations for the rest were partially helical and were generated from a 350-ns simulation of α -synuclein in water starting from the NMR structure (PDB ID 1XQ8) [64]. The protein was placed in explicit solvent (SPC/E water [65]) with 150 mM NaCl in all of the simulations. In the absence of direct interactions between the salt and α -synuclein, the conformational properties of α -synuclein depend only on the ionic strength of the solution. The ionic strength considered in the simulations was similar to the conditions considered in the measurements. We expect any variation in conformational properties of α -synuclein that may have caused by small differences in ionic strength will become insignificant during the Bayesian analysis. All simulations were performed using the constant pressure and constant temperature (NPT) ensemble with the OPLS-AA force field [66,67] in GROMACS [68]. The OPLS-AA force field was used for all simulations because the percent helicity was consistent with CD measurements of α -synuclein.

2.3. Simulated IR spectrum

The amide-I band was calculated from MD simulations, as was done in many previous studies [26,31,37,69,70]. The Hamiltonian matrix was constructed for each configuration to calculate the amide-I band of each cluster. The diagonal elements of the Hamiltonian matrix are sensitive to the environment of the amide group, which causes a shift in the fundamental frequency of each chromophore, while the off-diagonal elements refer to the coupling between different chromophores.

We utilized two models to calculate the shift in fundamental frequency of the chromophore (diagonal elements of the Hamiltonian):

- (1) *Hydrogen bond induced shift in frequency*: The shift in frequency for amide-I stretching was assumed to be a linear combination of all the hydrogen bonding interactions between the peptide carbonyl and hydrogen bond donors. Similar to other studies [26,69,70], a hydrogen bond is assumed to exist when the distance between the hydrogen of a donor and the oxygen of the peptide unit is less than the H-bond cutoff distance (2.6 Å) and makes favorable angles ($\text{C}=\text{O} \cdots \text{H}$ and $\text{O} \cdots \text{H}-\text{X} > 90^\circ$). When these constraints are satisfied, the diagonal shift in frequency ν_{H} due to H-bonding is given in cm^{-1} as:

$$\delta\nu_{\text{H}} = 30(r_{\text{OH}} - 2.6) \quad (1)$$

where r_{OH} is the $\text{C}=\text{O} \cdots \text{H}-\text{X}$ distance in Å. The factor of 30 is in quantitative agreement with *ab initio* calculations for structures near the equilibrium hydrogen bonding distances [71,72]. The hydrogen bond-dependent frequency shift map was estimated by simulating the amide-I band of N-methyl acetamide using *ab initio* molecular orbital calculations [71,72]. The fundamental frequency of N-methylacetamide was found to be sensitive to both the

dielectric effect and the formation of hydrogen bonds. The empirical model of hydrogen bond shift (given by Eq. (1)) accounts for both of those effects. It was also assumed that shifts in fundamental frequency due to solvation are additive in terms of the different hydrogen bonds that a particular peptide carbonyl group can form with water molecules. The unperturbed frequency of the oscillator was chosen as 1673 cm^{-1} based on reproducing the peak of the measured amide-I band of α -synuclein.

- (2) *Electric field induced shift in frequency:* The amide-I local-mode frequencies were calculated with the empirical frequency map developed in [33]:

$$\delta\nu_H = 5099E_{i,C} + 551E_{i,N} \quad (2)$$

In the map, the local-mode frequencies are related to the electric field due to the environment of the chromophore. $E_{i,C}$ is the electric field on the C atom, while $E_{i,N}$ is the electric field on the N atom of the i th amide group along the C=O direction (in atomic units). A $20\text{ \AA}$ cutoff is used for electric field calculations in the MD simulations. The electric field shift is only calculated for atoms more than three bonds from the chromophore in this model. The unperturbed frequency of the oscillator was chosen as 1700 cm^{-1} based on reproducing the peak of the measured amide-I band of α -synuclein. In both of the above models, the off-diagonal terms in the Hamiltonian matrix were calculated using an empirical dipole–dipole coupling model [13,73]. The coupling modifies the nearly degenerate amide-I modes and contributes to the off-diagonal terms of an exciton Hamiltonian matrix. The couplings between adjacent peptide units are mainly through-bond interactions, whereas those between nonadjacent peptide units are primarily long-distance electrostatic interactions. The couplings between neighboring residues depend on the backbone dihedral angles close to the corresponding chromophore, and were calculated using the approach of nearest-neighbor coupling map of Jansen et al. [30]. All other coupling terms were calculated using the transition dipole coupling approximation, β_{km} :

$$\beta_{km} = \frac{\vec{\mu}_k \cdot \vec{\mu}_m - 3(\vec{\eta}_{km} \cdot \vec{\mu}_m)}{r_{km}^3} \quad (3)$$

where $\vec{\mu}_k$ is the effective transition dipole of the k th amide-I mode, $\vec{\eta}_{km}$ is the unit vector connecting the dipoles, and r_{km} is the distance between the dipoles. The empirical transition dipole moment has a magnitude of 0.305 D and is located $0.868\text{ \AA}$ from the carbonyl carbon along the CO bond, directed 20° from the CO bond toward the N atom [13].

For each configuration from simulation trajectories, the Hamiltonian matrix was constructed and diagonalized to obtain the excitonic frequencies and intensities. A typical [70] homogeneous dephasing time (T_2) of 0.5 ps was used. The ensemble averaging of the frequency spectrum over all configurations obtained from the simulation incorporates the inhomogeneous contribution.

2.4. Clustering based on secondary structure and the backbone SASA of protein

We clustered the configurations generated using MD simulations based on the solvent exposure of backbone atoms and the helical content. Secondary structure analysis was performed using the program DSSP [74] in the GROMACS analysis suite [68]. DSSP identifies the hydrogen bonds in the protein by calculating the electrostatic interaction energy between hydrogen bond donors and acceptors. The helices in the protein at every time step were calculated as residues that were in an α -helix, a 3–10 helix, or a β -helix. The backbone solvation was measured by estimating the solvent accessible surface area (SASA). The SASA of the backbone atoms was calculated using the Shrake–Rupley algorithm in the

GROMACS analysis suite [68]. The protein was clustered based on backbone SASA and percent helicity in two dimensions. We considered ten partitions of equal width along the dimensions of SASA (2.8 nm²) and helicity (5.6%). A total of 67 clusters were formed with at least 100 conformations in each cluster based on these criteria. A maximum of 1000 conformations were chosen to represent each cluster, as the IR spectrum had converged within 1000 conformations (data not shown). For partitions with more than 1000 configurations, we ensured that a distinct ensemble of conformations represented the cluster in the IR calculations by combining configurations in different simulations (that fell within the dimensions of the cluster) and by ensuring that the most distant conformations in a simulation (by time) were chosen.

2.5. Bayesian fit to IR data

We employed a Bayesian formulation similar to the BW algorithm [41,42] to estimate the thermodynamic weights or probability ($\vec{w}$) of a protein to exist in each conformational cluster defined above, such that the amide-I band of the protein was consistent with the measured spectrum. The protein is assumed to exist in an ensemble of conformations belonging to each cluster with a certain probability. The intensity of the amide-I band for a particular set of weights $\vec{w}$ ($I_{\vec{w}} = \{I_{w(i)}\}$) where i represents the frequency in wave number and varies between 1600 and 1700 cm⁻¹) of the protein is:

$$I_{\vec{w}}(\omega_j) = \sum_{i=1}^N w_i I_i(\omega_j) \quad (4)$$

where w_i and $I_i(\omega_j)$ represent the thermodynamic weight (or the fraction of time spent in each cluster) and intensity of the IR spectrum of the i th cluster at ω_j frequency respectively, and $N=67$ refers to the total number of clusters. The measurement was performed at 105 different wavelengths, and j can vary between 1 and $M=105$. As w_i represents the thermodynamic weight of each cluster, the following conditions apply to these weights:

$$\sum_i w_i = 1 \text{ \& } w_i \geq 0 \text{ for all } i.$$

In the Bayesian approach, the probability of any given set of weights given the experimental data (posterior probability) was estimated using a Monte Carlo Markov chain simulation. The mean of the weights and the uncertainty in the fit (i.e., whether vastly different sets of weights can fit the data equally well) was calculated using the posterior probability distribution.

According to Bayes theorem [75], the posterior probability density ($f_{\vec{w}|\vec{I}_e}$) for the weights given the experimental data is:

$$f_{\vec{w}|\vec{I}_e}(\vec{w}|\vec{I}_e) = \frac{f_{\vec{I}_e|\vec{w}}(\vec{I}_e|\vec{w})f_{\vec{w}}(\vec{w})}{\int d\vec{w} f_{\vec{I}_e|\vec{w}}(\vec{I}_e|\vec{w})f_{\vec{w}}(\vec{w})} \quad (5)$$

where $f_{\vec{w}}(\vec{w})$ is the *a priori* probability of the weights of each cluster and $f_{\vec{I}_e|\vec{w}}(\vec{I}_e|\vec{w})$ represents the likelihood of obtaining the experimental IR spectrum $\vec{I}_e$ given the set of weights $\vec{w}$. The denominator is a constant that represents the probability of obtaining the experimental spectrum given any set of weights $f_{\vec{I}_e}$ and is typically not calculated during the Bayesian fit. The details of each term in the numerator in our model are given below.

Prior Distribution, $f_w(\vec{w})$ —The best priors are based on the difference in thermodynamic stability between different clusters of the protein. However, it is very difficult to measure the free energy differences between different states of the protein using MD simulations. In the absence of free energy estimates, the priors are chosen to be Gaussian distributions around means based on the potential energy (enthalpically driven priors) or entropy (entropically driven priors) for each cluster. In the enthalpically estimated priors, we take the center of the prior weight of each cluster based upon the Boltzmann distribution of the potential energy of the system:

$$w_i^P = \frac{\langle e^{-U(c_i)/k_B T} \rangle}{\sum_{j=1}^N \langle e^{-U(c_j)/k_B T} \rangle} \quad (6)$$

where the superscript “ P ” stands for the mean of the prior distribution, $U(c_i)$ represents the potential energy of a conformation c in cluster i , k_B represents the Boltzmann constant, and T represents the temperature of the system (298 K). The angular brackets represent the average value of the quantity within the brackets for all the conformations in the cluster. In addition to the enthalpically driven priors, we also calculated another set of weights from entropy by substituting $U(c_i)$ with $TS(c_i)$, where $S(c_i)$ is the entropy of the ensemble of the protein. The entropy was estimated using a simple polymer model [76].

As the weights represent the probabilities of the protein to be present in a particular cluster (i.e., $0 \leq w_i \leq 1$, and $\sum_{i=1}^N w_i = 1$), a simple Gaussian around the center of the prior weight is not ideal for the Bayesian calculation. Instead, in the BW algorithm [41,42], the N -dimensional weights are transformed into the $(N-1)$ -dimensional Euclidean space in which each coordinate h_i ($i = 1, \dots, N-1$) is given by the equation:

$$h_i = \frac{1}{\sqrt{i(i+1)}} \left(\sum_{j=1}^i \ln(w_j) - i \ln(w_{j+1}) \right) \quad (7)$$

The prior distribution was assumed to be a Gaussian in the $(N-1)$ -dimensional Euclidean space and is given by:

$$f_{\vec{w}|k}(\vec{w}|k) = \left(\prod_{i=1}^N w_i^{-1} \right)^{-1} \exp \left(- \sum_{i=1}^{N-1} k \frac{(h_i - h_i^P)^2}{2} \right) \quad (8)$$

where $\vec{h}^P = (h_1^P, \dots, h_{N-1}^P)$ represents the point in the $(N-1)$ -dimensional Euclidean space corresponding to $\vec{w}^P$ and $(\prod_{i=1}^N w_i^{-1})^{-1}$ is the Jacobian of the coordinate transformation. This represents the analogue of the Gaussian distribution for the vector of weights. We could have chosen the variance of the Gaussian distribution k^{-1} based on the variance of the Boltzmann distribution of each cluster. However, in the absence of precise free energy calculations, the prior distribution should take into account large uncertainties in the mean of the priors. Unfortunately, these uncertainties are not known *a priori*. Hence, we utilized a uniform mixture of Gaussians with different widths k so as to reduce the sensitivities of posterior probabilities to the prior weights $\vec{w}^P$. This was achieved by integrating the distribution for a uniform distribution for all positive values of k . In other words, the prior distribution is given by:

$$f_{\vec{w}}(\vec{w}) = \int_{k=0}^{k=\infty} f_{\vec{w}|k}(\vec{w}|k) f_k(k) dk \quad (9)$$

Likelihood function, $f_{\vec{w}}(\vec{w})$ —The likelihood function represents the probability of getting the experimental spectrum for a given set of weights while accommodating the errors in the measurement and theoretical prediction. It is independent of the prior distribution discussed above. The posterior distribution is the product of the likelihood function and the prior distribution. Similar to the BW algorithm [41,42], the likelihood of producing the experimental spectrum for a given set of weights is approximated using a Gaussian model:

$$f_{\vec{w}}(\vec{w}) = [2\pi(\varepsilon_{\text{IR}}^2 + \alpha_{\text{IR}}^2)]^{-0.5} \exp \left[\frac{-\sum_{j=1}^M (I_e(\omega_j) - \sum_{i=1}^N w_i I_i(\omega_j))^2}{2(\varepsilon_{\text{IR}}^2 + \alpha_{\text{IR}}^2)} \right] \quad (10)$$

where $\varepsilon_{\text{IR}}^2$ and α_{IR}^2 are the experimental and the prediction errors in the IR spectrum, respectively. It should be noted that Eq. (10) represents an approximation to the true likelihood, which can, in principle, be calculated exactly but it depends on the details of the error model. The numerator in the Gaussian term represents the difference between the experimental spectrum and the predicted spectrum (with a particular set of weights) at a single frequency. The intensity of the IR spectrum at each wave number is assumed to be independent of the intensity at all other points while calculating the IR spectrum. It should be noted that with this assumption we neglect any correlated changes in intensity at different points due to correlated “error”, either experimental or theoretical.

We performed the Bayesian fit to the data under three conditions: (a) low error – similar to a least square fit (ε_{IR} and $\alpha_{\text{IR}} = 0.001$ units or 2% error at peak of the experimental spectrum), (b) medium error (ε_{IR} and $\alpha_{\text{IR}} = 0.002$ units or 6% error at peak of the experimental spectrum), and (c) large error (ε_{IR} and $\alpha_{\text{IR}} = 0.01$ units or 20% error at peak of the experimental spectrum). The expected mean square deviation between the experimental and computed spectra is the sum of the squares of experimental and computational errors. In this model, the difference between the experimental and computed spectra at each frequency and the total expected error determine the likelihood for producing the experimental spectrum with a particular set of weights. For simplicity, we assume that the error has a Gaussian distribution and is constant over the whole frequency range in this manuscript. A frequency dependent error can be utilized to examine the regions in the spectrum that lead to resolution in weights for different clusters.

2.5.1. Monte Carlo algorithm—A Markov chain Monte Carlo (MCMC) algorithm was used to calculate the posterior probability distribution of the weights given the data [77–79]. The posterior density was iteratively used to sample weights from their conditional distributions. Gibbs sampling [80] was performed over k , which was then used to estimate the prior probability at each step. A Metropolis–Hastings step [78] was implemented for sampling the weights using a simple normal distribution centered at the current weight vector as the proposed distribution. The proposed distribution had an isotropic variance that was tuned during an equilibration period so that about 25% of all steps were accepted. Similar to the BW algorithm, to ensure improved sampling of the posterior distribution, a multiple-replica approach was employed, with simulations performed at different “temperatures” [81,82]. We used 10 different replicas varying between the temperatures of 1 and 1.5, and the percentage of moves accepted between different temperatures was adjusted to 30%. The MCMC was initially equilibrated for 1 million steps to remove any memory of the initial conditions. Swaps were accepted every 100 steps, and the production runs were

performed for 10^7 steps. The posterior distribution was sampled at 10,000 different weight vectors. We also performed the simulations 10 times to evaluate the convergence of the parameters. The value of the mean Bayesian weights and uncertainty reported in the Results section were calculated over all ten simulations. These values were also compared to those obtained using individual simulations.

2.5.2. Analysis of the posterior distribution—The average weight for each cluster is:

$$w_j^B = \langle w_j^B \rangle = \int d\vec{w} w_j f_{\vec{w}|\vec{I}_e}(\vec{w}|\vec{I}_e) \quad (11)$$

where w^B is the Bayesian estimate for the mean weight of each cluster. If two different sets of weights have highly posterior probability, the posterior probability can be considered to be highly redundant. We utilized the expected value of the Jensen–Shannon divergence (JSD) to measure the redundancy of the Bayesian fit. The measure of difference between two sets of weight vectors is based on the Jensen–Shannon divergence:

$$\Omega^2(\vec{w}^a, \vec{w}^b) = S\left(\frac{\vec{w}^a + \vec{w}^b}{2}\right) - \frac{1}{2}(S(\vec{w}^a) + S(\vec{w}^b)) \quad (12)$$

where $S(\vec{w}^a) = -\sum_{i=1}^N w_i \ln(w_i)$ is the information entropy [83]. This quantity has the property that $0 \leq \Omega^2(\vec{w}^a, \vec{w}^b) \leq 1$ and it is equal to 0 only if $\vec{w}^a = \vec{w}^b$. The average value of JSD in the posterior distribution is used to measure the posterior expected divergence:

$$\sigma_{\vec{w}^B} = \langle \Omega^2(\vec{w}^B, \vec{w}) \rangle_{BW}^{0.5} = \left[\int d\vec{w} \Omega^2(\vec{w}^B, \vec{w}) f_{\vec{w}|\vec{I}_e}(\vec{w}|\vec{I}_e) \right]^{0.5} \quad (13)$$

The uncertainty in the fit is measured by $\sigma_{\vec{w}^B}$ (which is similar to standard deviation of fit), and its value is such that $0 \leq \sigma_{\vec{w}^B} \leq 1$. It is zero only if there is no uncertainty in thermodynamic weights.

2.5.3. Root mean square deviation of fit—The root mean square deviation of the fit is a measure of the goodness of the fit and is measured using the formula:

$$\text{RMSD} = \sqrt{\frac{\sum_{j=1}^M (I_e(\omega_j) - I_{\vec{w}^B}(\omega_j))^2}{M}} \quad (14)$$

The RMSDs for the prediction of the chemical shifts were also calculated similarly.

2.6. Chemical shift

The chemical shift for each frame in a cluster was calculated using the program SHIFTX [84], which utilizes a semi-empirical approach to calculate the effect of the environment on the chemical shift of each atom. As the conformations within a cluster were distinct and were chosen from different simulations when possible, the chemical shift of any atom in a given cluster was taken as the mean chemical shift for that atom in all frames within that particular cluster. The ensemble chemical shift was calculated as the sum of the chemical shifts for each cluster multiplied by the thermodynamic weight of that cluster.

3. Results and discussion

3.1. Experimental spectrum

The IR amide-I' band (the prime denotes the frequencies of the deuterated amide group) is sensitive to the conformation and solvation of the polypeptide backbone and therefore reports on secondary structure formation. The FT-IR spectrum of α -synuclein at 25 °C has a broad envelope centered near 1645 cm^{-1} due to the predominantly disordered conformation of the polypeptide backbone (Fig. 1). Due to considerable overlap of the structural components in the raw absorbance spectra, inferring the percentage of different secondary structure elements from the IR spectrum in the amide-I' band of α -synuclein is difficult. Our peak fitting analysis reveals six major peaks in the amide-I' band at 1615, 1630, 1645, 1657, 1672, and 1685 cm^{-1} and we made tentative secondary structure assignments to each of these peaks [21,85–88]. The area under the curve corresponding to each peak can be used to calculate the percentage of different secondary structure elements in the protein. The 1615 and 1685 cm^{-1} peaks are assigned to β -sheet structures, which correspond to about 6% of the absorbance area. The peak at 1630 cm^{-1} is assigned to β -sheet structures and solvated helix, and the content is about 19%. The peak at 1645 cm^{-1} is assigned to disordered structures and constitutes about 33% of the total area. The peak at 1672 cm^{-1} is about 17% and is assigned to turn structures, while the peak at 1657 cm^{-1} is about 25% and was assigned to helical structures. While the peak fitting above can be a reasonable approach to extract secondary structure contents of well-folded proteins, we believe that it can lead to inaccurate estimates for IDPs. For example, CD measurements have shown that the helical content of α -synuclein is about 5–10% [89]. However, with this peak fitting routine, the helical content appears to be overestimated. Therefore new fitting methods need to be developed to extract the secondary structure contents of α -synuclein and perhaps other IDPs from the amide-I band.

Below, we use the experimental spectrum of α -synuclein as a basis to illustrate how a Bayesian framework can be used to deduce thermodynamic weights of different conformational states contributing to this observed spectrum. We generated conformations from 100 different MD simulations and clustered them according to chosen physical properties. Since the thermodynamic weights of these conformational clusters remain unknown at the end of the MD simulations, we used a Bayesian framework to calculate the weights of these clusters, such that the overall amide-I band of the α -synuclein ensemble was most consistent with the observed spectrum. We utilized this approach to make estimates of the properties of the α -synuclein such as size, secondary structure content, and tertiary contacts.

3.2. Influence of compactness and residual secondary structure of IDPs on the amide-I band

First, we explored the conformational properties of α -synuclein that influence the amide-I band. These properties are eventually used to cluster conformations for the Bayesian approach. We performed more than 100 MD simulations of α -synuclein in explicit solvent. The protein formed a heterogeneous compact state that was stable up to hundreds of nanoseconds due to the formation of long-range tertiary contacts within the protein. α -synuclein forms a heterogeneous ensemble of conformations in our MD simulations, with sizes and helicity varying between 20 and 65 Å and 0% and 52%, respectively [39]. The amide-I band is sensitive to the presence of secondary structural elements in the protein due to the difference in the pattern of intra-molecular and inter-molecular hydrogen bond patterns of the amide group [39]. However, the pattern of intra-molecular and inter-molecular hydrogen bond patterns can also vary due to tertiary interactions in the protein. As the protein becomes more compact, a number of solvent–protein hydrogen bonds are

replaced by intra-molecular hydrogen bonds. Therefore, one should be able to detect and distinguish signatures in the amide-I band that arise due to variation in the size of an IDP.

Here, we evaluated how the compactness of an IDP could impact the amide-I band, along with any residual secondary structural elements present. In order to perform this analysis, we simulated the amide-I band of four conformational clusters of α -synuclein that varied either in their number of intramolecular contacts (measured by backbone SASA) and/or percentage helicity (Fig. 2). These clusters are named extended coil ($\approx 89 \text{ nm}^2$ SASA and $\approx 3\%$ helical), collapsed coil ($\approx 81 \text{ nm}^2$ SASA and $\approx 3\%$ helical), extended helical ($\approx 89 \text{ nm}^2$ SASA, $\approx 32\%$ helical), and collapsed helical ($\approx 81 \text{ nm}^2$ SASA, $\approx 32\%$ helical). The amide-I band of these clusters was calculated using two different empirical approaches. In the first model, every hydrogen bond between a carbonyl group and its environment was assumed to cause a shift in the fundamental frequency of the amide group [70]. In the second model, the fundamental frequency shift was calculated as a linear combination of the electric field on the carbon and nitrogen atoms of the corresponding amide group [33]. The dipole-dipole coupling model [13] was used to calculate the off-diagonal terms for non-neighboring residues of the Hamiltonian matrix, while the backbone torsion angles were utilized to calculate the nearest neighbor coupling terms in both models [30].

The calculated amide-I bands of all four conformational clusters are very similar to each other in the first model, as seen in Fig. 3. Also, all bands are similar to the experimentally determined overall amide-I band of α -synuclein in water. However, there are subtle differences between the bands simulated for each cluster. As expected, the presence of helicity in the protein causes a red shift (decrease in wavenumber) in both the compact and extended conformations. This red shift is due to an increase in the coupling between neighboring chromophores in partially helical conformations and to an increase in intra-molecular hydrogen bonds between backbone atoms. This shift is larger in the extended state (3 cm^{-1}) as compared to the collapsed state (1.2 cm^{-1}). On comparing the coil states with the corresponding partially helical ones, all backbone hydrogen bonds in the partially helical state are replaced by hydrogen bonds with the solvent or sidechains in the coil, resulting in nearly similar hydrogen bond-induced shifts between the partially helical and random coil states (Fig. 4). In addition, the expanded conformations display a larger width in the amide-I band than the collapsed conformation due to the wide distribution in the hydrogen bond-induced shifts in these states (Fig. 3). In addition, the coupling terms cause a broadening and a red-shift of the partially helical states (Figure S1). Thus, a combination of hydrogen bond-induced frequency shifts and coupling terms causes subtle changes in the amide-I band of each cluster.

The amide-I bands of the four conformational clusters were also calculated using the second model, which uses transferable electric field-based IR frequency shifts. We obtain similar qualitative shifts in peak and width of spectrum for the four clusters (Figure S2). However, the widths and peaks of the amide-I bands are much larger than the width of measured amide-I band due to larger variations in the solvent-induced electric field on the carbonyl group (Figure S3). This is consistent with the simulated amide-I band of rat amylin peptides which was found to be broader than the measured spectrum [31]. In the rest of the paper, we simulate the amide-I band for each cluster using the empirical hydrogen bond induced shift for each cluster since that is more similar in shape to the experimental spectrum than the spectrum generated from the electric field-induced shift method.

3.3. Clustering of conformations from MD based on compactness and secondary structure

Based on past studies on α -synuclein [39,42,46,48,49,54, 59,60], we expect this IDP to exist in an ensemble of buried and solvent exposed states in aqueous environment. We split the configurations generated during the MD simulation based on a combination of amount of

solvent accessibility and helical content. The backbone SASA of the protein varied between 100 and 72 nm² while the helical content of the protein varies between 0% and 56%. We generate 67 clusters of the protein, such that at least 100 conformations are present in each cluster (Table 1). A maximum of 1000 distinct conformations are chosen for the IR calculations in each cluster. We visually inspect the IR spectra to ensure that the simulated spectra for each cluster were smooth and did not contain excessive noise. The experimental IR spectrum is a linear combination of the IR spectrum of these states of the protein, weighted by the thermodynamic contribution of each state.

3.4. Predicting thermodynamic weights of conformational clusters using IR spectroscopy

Given the experimental IR spectrum of α -synuclein, we utilize a Bayesian approach similar to the BW algorithm [41,42] to estimate the thermodynamic weight of all 67 conformational clusters defined above. In the BW method, the probability of the weights given the experimental spectrum (posterior probability) is proportional to the product of two probability terms as shown in Eq. (5). Details of this methodology are provided in the methods section. The first term, $f_{\mathcal{D}|\mathbf{w}}(\mathcal{D}|\mathbf{w})$, is the likelihood of obtaining the spectrum given the weights, and is obtained by calculating the amount of mismatch between the experimental and simulated spectrum for a particular set of weights. The Bayesian approach accounts for errors in the experimental measurement as well as the theoretical estimation of the IR spectrum while calculating the likelihood function. The second term, $f_{\mathbf{w}}(\mathbf{w})$, is the prior probability of the cluster. Here, we consider two such probabilities: the Boltzmann distribution of the potential energy and entropy terms. The posterior probabilities can then be used to calculate the mean weight of each cluster, and to estimate the redundancy or amount of uncertainty in the fit. The BW method has already been tested and applied to study ensembles of moderately sized IDPs given the proteins' NMR chemical shifts and residual dipolar couplings [41,42]. In this manuscript, we extend the BW method to infer ensemble properties of α -synuclein in an aqueous environment given its IR spectrum.

We apply this method under three sets of conditions assuming a low, medium, or large amount of error in the IR data and/or simulation method (see methods). If errors from measurements and calculated IR spectra are very low (~ 1 – 2 cm⁻¹ in peak position), then the BW algorithm can be used to estimate the weights that give the best fit to measurements. In this limit, the BW algorithm provides precise information about the weights of each cluster (Fig. 5B), and the simulated spectrum agrees reasonably with the measured spectrum (Fig. 5A). The weights for all 100 clusters are shown in Fig. 5B, but only 67 clusters are highly populated in the MD simulations and are used in the Bayesian fit. The amount of uncertainty in the fit is measured by the expected divergence ($\mathcal{D}|\mathbf{w}$) as shown in Eq. (13), which varies between 0 and 1. An expected divergence of 0 implies that the predicted weights for each cluster are highly unique, while an expected divergence of 1 implies that the weights for each cluster are highly redundant. Under the low error limit, the algorithm predicts that the protein is either present in a collapsed helical state or in an extended coil conformation. The expected divergence is $\mathcal{D}|\mathbf{w} = 0.008$ for this fit. In addition, we verify that the fit is unique by performing ten MCMC simulations with different random seeds to obtain highly similar Bayesian weights. Thus, the Bayesian framework provides a reasonable, unique fit to measurements and also can be used to extract weights from clusters with varying SASA and secondary structure when the experimental measurements and theoretical predictions contain very small errors.

As expected, the uniqueness of the fit relies on the magnitude of the errors that arise from measurements and predictions. In the case of a high error (~ 10 – 20 cm⁻¹ in peak position), the prior probability distribution, $f_{\mathbf{w}}(\mathbf{w})$, dominates the weights estimated using the BW algorithm (Fig. 5C and D). Since the high error masks the distinction between the spectra provided by different sets of allowable weights, the measured spectrum cannot be used

reliably to provide ensemble properties of the system. The divergence is $wB = 0.13$, which indicates that the fit is redundant. However, the Bayesian fits converged, as ten different MCMC simulations with different initial random seeds provided similar Bayesian weights. We further confirm that Bayesian weights in the high error limit are sensitive to the mean of the prior distribution by carrying out additional calculations with priors determined by the entropy of different conformational clusters (Figure S5).

A more appropriate error estimate for calculating the likelihood function is based on the root mean square deviation of the Bayesian or least square fits to the experimental measurements. The root mean square deviation between the experimental measurement and the theoretical prediction for the amide-I band is in the range of 0.003 units (low error limit) and 0.004 units (high error limit). This motivated us to run the BW algorithm under the condition that the expected deviation between the theoretical and experimental spectra is close to 0.004 units (medium error limit corresponding to differences in peak of $\approx 4\text{--}5\text{ cm}^{-1}$). In the medium error limit, the posterior distribution is calculated as a balance between obtaining a fit to the experimental spectrum and the prior probability distribution (Fig. 5E and F). In this limit, the fit is as redundant as in the high error limit $wB = 0.14$, while the Bayesian weights are very different from the prior weights, indicating that the spectrum does provide additional information that leads to a better fit with the experimental spectrum (root mean square deviation = 0.0033 units). The Bayesian fit converges, and the weights are shown in Fig. 5F and S4 for all 10 simulations.

The Bayesian weights suggest that 65% of the time the backbone of α -synuclein is solvent exposed with minimal residual secondary structure content. About 56% of the contribution to the amide-I band originates from a conformational ensemble with backbone SASA of $92.4\text{--}95.2\text{ nm}^2$ and helicity of $0\text{--}5.2\%$. The second dominant contribution (8.2%) is from a conformational ensemble with backbone SASA in the range of $89.6\text{--}92.4\text{ nm}^2$ and helicity of $5.2\text{--}10.4\%$. The mean helicity in the Bayesian ensemble is 8.9%, which is consistent with CD measurements that show that α -synuclein has $5\text{--}10\%$ residual secondary structure [89]. On performing the Bayesian analysis of the amide-I band after removing the two most dominant clusters, the fit displays larger error (root mean square deviation = 0.0038 units) and redundancy $wB = 0.19$. In addition, the estimated weight of the neighboring cluster (in terms of SASA and helicity) shows an increase, thereby providing further evidence that α -synuclein might spend a large fraction of the time in conformations belonging to those two clusters that were removed (Figure S6). In principle, the process of error estimation should be independent from application of the method.

3.5. Validation – ensemble properties based on predicted thermodynamic weights

In a rigorous test of this BW technique, we utilize the above thermodynamic weights to compare the theoretically expected chemical shift of hydrogen and carbon atoms in the amide group of each residue with the corresponding experimentally determined quantities [50]. We use SHIFTX [84] to calculate these chemical shifts for each conformation of the protein in the different clusters. The chemical shift for hydrogen and carbonyl atoms in a cluster is assumed to be the mean of the chemical shifts of the corresponding atoms from all conformations that belonged to that cluster. This assumption is consistent with the Bayesian framework discussed above. This framework estimates the thermodynamic weights for each cluster, and it does not distinguish between different conformations within a cluster. Also, the posterior distributions are very different from the prior distributions, indicating that potential energy or entropy estimates of different conformations cannot be reliably used to distinguish the weights of different conformations within a cluster. The predicted chemical shift for a particular atom in the Bayesian ensemble of α -synuclein is then calculated as the weighted average of the chemical shift from all clusters.

As depicted in Fig. 6 and Table 2, the predicted chemical shifts of all backbone atoms are consistently similar to measurements in the ensembles where Bayesian fits were carried out under low and medium error limits. The chemical shifts for the amide hydrogen and carbonyl carbon atoms in the amide group in the ensemble inferred assuming low errors display a correlation coefficient of 0.48 and 0.80, respectively. The corresponding root mean square deviations (RMSD) are 0.26 and 1.04 ppm compared to measurements. However, the best match is obtained with the thermodynamic weights estimated with medium error, which displayed a correlation coefficient of 0.58 and 0.89 with the measurements of hydrogen and carbonyl carbon atoms in the amide group, respectively. The corresponding RMSDs are 0.23 and 1.16 ppm. These results imply that the Bayesian fit obtained using low error (similar to least squares fit) might be over fitting the amide-I band. The differences between the ensemble chemical shift in the medium error limit and the measured quantities are within the accuracy limit of SHIFTX based on comparison of predicted and measured chemical shifts for folded proteins (Correlation coefficient of 0.58 and 0.863, and RMSD of 0.232 and 1.16 ppm for hydrogen and carbonyl carbon atoms of the amide groups respectively) [84]. In addition, the predicted chemical shifts for nitrogen atom of the amide group and backbone C α atom are in agreement with measurements (Table 2).

In order to emphasize that information from the amide-I band of α -synuclein led to better prediction of the chemical shift, we also compare the chemical shifts predicted for the high error Bayesian ensembles (both enthalpy and entropy-based priors) to the experimentally determined chemical shifts. In the high error limit, the Bayesian fits do not utilize as much information from the IR spectrum, and the predicted weights are closer to the mean of the prior distribution that were estimated using the potential energy or entropy (Table 2 and Fig. 6). The chemical shifts for the amide groups using these ensembles show a larger RMSD (0.43 and 0.41 ppm for hydrogen atoms; 1.40 and 1.25 ppm for carbonyl carbon atoms in the amide group for enthalpically and entropically determined priors, respectively) and display much lower correlation (0.29 and 0.30 for hydrogen atoms; 0.57 and 0.64 for carbonyl carbon atoms for enthalpically and entropically determined priors, respectively). Even though the chemical shifts for the nitrogen and C α atoms are less sensitive to different ensembles, predicted chemical shifts for them using the ensembles with low and medium error limits are in slightly better agreement with measurements. Therefore, high error Bayesian ensembles display much lower correlation to the experimentally determined chemical shifts than the low and medium error Bayesian ensembles (Table 2). In other words, information inferred using the amide-I band does improve the prediction of the chemical shifts of backbone atoms.

4. Conclusions

It is difficult to characterize the conformational variability of IDPs both experimentally and computationally. The amide-I band contains information on the solvation and residual secondary structure contents of IDPs, but it is challenging to infer these ensemble properties. We develop a Bayesian framework to infer such properties of α -synuclein using a combination of MD simulations and the experimentally measured amide-I band. We use the MD simulations to generate several distinct conformational clusters with varying level of solvent exposure and secondary structure content. We employ Bayesian formulation to estimate the thermodynamic weights of these clusters that are consistent with the experimental amide-I band and non-redundant. These thermodynamic weights are representative of the fraction of time that α -synuclein spends within each distinct ensemble of conformations during measurements. The results from the Bayesian framework analysis indicate that α -synuclein is mostly solvent-exposed and contains minimal secondary structure about 60% of the time. In a rigorous test of this approach, we show that the chemical shifts of α -synuclein predicted from Bayesian analysis based thermodynamic

weights that satisfy the amide-I band are more consistent with measured chemical shifts than those that were constructed in the absence of such information.

In summary, we propose a Bayesian framework to estimate the thermodynamic weights of heterogeneous conformational ensembles of relatively large intrinsically disordered proteins from experimentally determined infrared spectra. We believe such a framework can be used for many different purposes. For example, a practical application of the proposed approach could be a refinement of force fields for MD simulations of IDPs. Enhanced sampling methods such as replica exchange MD simulations of smaller IDPs (<50 residues) can be used to directly obtain the thermodynamic weights of distinct conformational populations. Then, these weights can be quantitatively compared to those deduced from measured IR spectra using a Bayesian framework for refinement of force fields.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by a LANL/LDRD Grant, NIH Grant R37-GM035556, and LANL Institutional Computing for the supercomputer time. A.S. was partially supported by a postdoctoral fellowship from the Center for Nonlinear Studies. We would like to thank Jennifer Mackie for editing. DMV would like to thank Timothy Causgrove for help with the experimental IR data analysis.

References

1. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Higgs KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. *J. Mol. Graph Model.* 2001; 19:26. [PubMed: 11381529]
2. Dunker AK, Silman I, Uversky VN, Sussman JL. *Curr. Opin. Struct. Biol.* 2008; 18:756. [PubMed: 18952168]
3. Liu J, Faeder JR, Camacho CJ. *Proc. Natl. Acad. Sci. USA.* 2009; 106:19819. [PubMed: 19903882]
4. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. *J. Mol. Biol.* 2004; 337:635. [PubMed: 15019783]
5. Papoian GA. *Proc. Natl. Acad. Sci. USA.* 2008; 105:14237. [PubMed: 18799750]
6. Tompa P. *Trends Biochem. Sci.* 2002; 27:527. [PubMed: 12368089]
7. Sethi A, Goldstein B, Gnanakaran S. *PLoS Comput. Biol.* 2011; 7:e1002192. [PubMed: 22022247]
8. Eliezer D. *Curr. Opin. Struct. Biol.* 2009; 19:23. [PubMed: 19162471]
9. Dyson HJ, Wright PE. *Chem. Rev.* 2004; 104:3607. [PubMed: 15303830]
10. Mittag T, Forman-Kay JD. *Curr. Opin. Struct. Biol.* 2007; 17:3. [PubMed: 17250999]
11. Kiefhaber T, Bachmann A, Jensen KS. *Curr. Opin. Struct. Biol.* 2012; 22:21. [PubMed: 22129832]
12. Barth A, Zscherp C. *Rev Q. Biophys.* 2002; 35:369.
13. Krimm S, Bandekar J. *Adv. Protein Chem.* 1986; 38:181. [PubMed: 3541539]
14. Arrondo JL, Muga A, Castresana J, Goni FM. *Prog. Biophys. Mol. Biol.* 1993; 59:23. [PubMed: 8419985]
15. Goormaghtigh E, Cabiaux V, Ruysschaert JM. *Subcell Biochem.* 1994; 23:405. [PubMed: 7855879]
16. Bour P, Keiderling TA. *J. Phys. Chem. B.* 2005; 109:23687. [PubMed: 16375349]
17. Decatur SM. *Acc. Chem. Res.* 2006; 39:169. [PubMed: 16548505]
18. Lim M, Hamm P, Hochstrasser RM. *Proc. Natl. Acad. Sci. USA.* 1998; 95:15315. [PubMed: 9860966]
19. Schweitzer-Stenner R. *Biophys. J.* 2002; 83:523. [PubMed: 12080139]

20. Natalello A, Ami D, Doglia SM. *Methods Mol. Biol.* 2012; 895:229. [PubMed: 22760323]
21. Susi H, Byler DM. *Methods Enzymol.* 1986; 130:290. [PubMed: 3773736]
22. Kong J, Yu S. *Acta Biochim. Biophys. Sin. (Shanghai).* 2007; 39:549. [PubMed: 17687489]
23. Byler DM, Susi H. *Biopolymers.* 1986; 25:469. [PubMed: 3697478]
24. Dong A, Huang P, Caughey WS. *Biochemistry.* 1990; 29:3303. [PubMed: 2159334]
25. Susi H, Byler DM. *Biochem. Biophys. Res. Commun.* 1983; 115:391. [PubMed: 6615537]
26. Brewer SH, Tang Y, Vu DM, Gnanakaran S, Raleigh DP, Dyer RB. *Biochemistry.* 2012; 51:5293. [PubMed: 22680405]
27. Jeon J, Yang S, Choi JH, Cho M. *Acc. Chem. Res.* 2009; 42:1280. [PubMed: 19456096]
28. Hayashi T, Zhuang W, Mukamel S. *J. Phys. Chem. A.* 2005; 109:9747. [PubMed: 16833288]
29. Zhuang W, Sgourakis NG, Li Z, Garcia AE, Mukamel S. *Proc. Natl. Acad. Sci. USA.* 2010; 107:15687. [PubMed: 20798063]
30. Jansen TIC, Dijkstra AG, Watsom TM, Hirst JD, Knoester J. *J. Chem. Phys.* 2006; 124:044502. [PubMed: 16460180]
31. Reddy AS, Wang L, Lin YS, Ling Y, Chopra M, Zanni MT, Skinner JL, De Pablo JJ. *Biophys. J.* 2010; 98:443. [PubMed: 20141758]
32. Ganim Z, Chung HS, Smith AW, Deflores LP, Jones KC, Tokmakoff A. *Acc. Chem. Res.* 2008; 41:432. [PubMed: 18288813]
33. Lin YS, Shorb JM, Mukherjee P, Zanni MT, Skinner JL. *J. Phys. Chem. B.* 2009; 113:592. [PubMed: 19053670]
34. Hayashi T, Mukamel S. *J. Phys. Chem. B.* 2007; 111:11032. [PubMed: 17725341]
35. Ham S, Kim J-H, Lee H, Cho M. *J. Chem. Phys.* 2003; 118:3491.
36. Maekawa H, Ge NH. *J. Phys. Chem. B.* 2010; 114:1434. [PubMed: 20050636]
37. Schmidt JR, Corcelli SA, Skinner JL. *J. Chem. Phys.* 2004; 121:8887. [PubMed: 15527353]
38. Rauscher S, Pomes R. *Biochem. Cell Biol.* 2010; 88:269. [PubMed: 20453929]
39. Sethi A, Tian J, Vu DM, Gnanakaran S. *Biophys. J.* 2012; 103:748. [PubMed: 22947936]
40. Lindorff-Larsen K, Trbovic N, Maragakis P, Piana S, Shaw DE. *J. Am. Chem. Soc.* 2012; 134:3787. [PubMed: 22339051]
41. Fisher CK, Huang A, Stultz CM. *J. Am. Chem. Soc.* 2010; 132:14919. [PubMed: 20925316]
42. Ullman O, Fisher CK, Stultz CM. *J. Am. Chem. Soc.* 2011; 133:19536. [PubMed: 22029383]
43. Dawson TM, Dawson VL. *Science.* 2003; 302:819. [PubMed: 14593166]
44. Auluck PK, Caraveo G, Lindquist S. *Annu. Rev. Cell Dev. Biol.* 2010; 26:211. [PubMed: 20500090]
45. Eliezer D, Kutluay E, Bussell R Jr, Browne G. *J. Mol. Biol.* 2001; 307:1061. [PubMed: 11286556]
46. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM. *J. Am. Chem. Soc.* 2005; 127:476. [PubMed: 15643843]
47. Bussell R Jr, Eliezer D. *J. Mol. Biol.* 2003; 329:763. [PubMed: 12787676]
48. Bernado P, Bertocini CW, Griesinger C, Zweckstetter M, Blackledge M. *J. Am. Chem. Soc.* 2005; 127:17968. [PubMed: 16366524]
49. Morar AS, Olteanu A, Young GB, Pielak GJ. *Protein Sci.* 2001; 10:2195. [PubMed: 11604526]
50. Rao JN, Kim YE, Park LS, Ulmer TS. *J. Mol. Biol.* 2009; 390:516. [PubMed: 19481090]
51. Jao CC, Der-Sarkissian A, Chen J, Langen R. *Proc. Natl. Acad. Sci. USA.* 2004; 101:8331. [PubMed: 15155902]
52. Drescher M, Veldhuis G, van Rooijen BD, Milikisyants S, Subramaniam V, Huber M. *J. Am. Chem. Soc.* 2008; 130:7796. [PubMed: 18512917]
53. Bortolus M, Tombolato F, Tessari I, Bisaglia M, Mammi S, Bubacco L, Ferrarini A, Maniero AL. *J. Am. Chem. Soc.* 2008; 130:6690. [PubMed: 18457394]
54. Lee JC, Langen R, Hummel PA, Gray HB, Winkler JR. *Proc. Natl. Acad. Sci. USA.* 2004; 101:16466. [PubMed: 15536128]
55. Chandra S, Chen X, Rizo J, Jahn R, Sudhof TC. *J. Biol. Chem.* 2003; 278:15313. [PubMed: 12586824]

56. Ferreon AC, Deniz AA. *Biochemistry*. 2007; 46:4499. [PubMed: 17378587]
57. Bussell R, Eliezer D Jr. *Biochemistry*. 2004; 43:4810. [PubMed: 15096050]
58. Perrin RJ, Woods WS, Clayton DF, George JM. *J. Biol. Chem.* 2000; 275:34393. [PubMed: 10952980]
59. Sandal M, Valle F, Tessari I, Mammi S, Bergantino E, Musiani F, Brucale M, Bubacco L, Samori B. *PLoS Biol.* 2008; 6:e6. [PubMed: 18198943]
60. Ferreon AC, Gambin Y, Lemke EA, Deniz AA. *Proc. Natl. Acad. Sci. USA.* 2009; 106:5645. [PubMed: 19293380]
61. Georgieva ER, Ramlall TF, Borbat PP, Freed JH, Eliezer D. *J. Am. Chem. Soc.* 2008; 130:12856. [PubMed: 18774805]
62. Tian J, Sethi A, Anunciado D, Vu DM, Gnanakaran S. *J. Phys. Chem. B.* 2012; 116:4417. [PubMed: 22439820]
63. Jha AK, Colubri A, Freed KF, Sosnick TR. *Proc. Natl. Acad. Sci. USA.* 2005; 102:13099. [PubMed: 16131545]
64. Ulmer TS, Bax A, Cole NB, Nussbaum RL. *J. Biol. Chem.* 2005; 280:9595. [PubMed: 15615727]
65. Jorgensen WL, Chandrashekar J, Madura JD, Impey RW, Klein ML. *J. Chem. Phys.* 1983; 79:926.
66. Jorgensen WL, Maxwell DS, Tirado-Rives J. *J. Am. Chem. Soc.* 1996; 118:11225.
67. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. *J. Phys. Chem. B.* 2001; 105:6474.
68. Hess B, Kutzner C, van der Spoel D, Lindahl E. *J. Chem. Theory Comput.* 2008; 4:435.
69. Gnanakaran S, Hochstrasser RM. *J. Am. Chem. Soc.* 2001; 123:12886. [PubMed: 11749547]
70. Gnanakaran S, Hochstrasser RM, Garcia AE. *Proc. Natl. Acad. Sci. USA.* 2004; 101:9229. [PubMed: 15197256]
71. Torii H, Tatsumi T, Tasumi M. *J. Raman Res.* 1999; 29:537.
72. Torii H, Tatsumi T, Kanazawa T, Tasumi M. *J. Phys. Chem. B.* 1998; 102:309.
73. Torii H, Tasumi M. *J. Chem. Phys.* 1992; 96:3379.
74. Kabsch W, Sander C. *Biopolymers.* 1983; 22:2577. [PubMed: 6667333]
75. Papoulis, A.; Pillai, U. *Probability. Mcgraw-Hill: Random Variables and Stochastic Processes;* 2002.
76. Zhou H-X, Dill KA. *Biochemistry.* 2001; 40:11289. [PubMed: 11560476]
77. Chib S, Greenberg E. *The Am. Statist.* 1995; 49:327.
78. Hastings W. *Biometrika.* 1970; 57:97.
79. Metropolis N, Ulam S. *J. Am. Statist. Soc.* 1949; 44:335.
80. Gelfand AE, Smith AFM, Amer J. *Stat. Ass.* 1990; 85:398.
81. Sugita Y, Okamoto Y. *Chem. Phys. Lett.* 1990; 314:141.
82. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcia AE. *Curr. Opin. Struct. Biol.* 2003; 13:168. [PubMed: 12727509]
83. Cover, T.; Thomas, J. *Elements of Information Theory.* Wiley Interscience; 2006.
84. Neal S, Nip AM, Zhang H, Wishart DS, *Biomol J. NMR.* 2003; 26:215.
85. Kazakov AS, Markov DI, Gusev NB, Levitsky DI. *Biophys. Chem.* 2009; 145:79. [PubMed: 19783089]
86. Surewicz WK, Mantsch HH. *Biochim. Biophys. Acta.* 1988; 952:115. [PubMed: 3276352]
87. Uversky VN, Li J, Fink AL. *J. Biol. Chem.* 2001; 276:10737. [PubMed: 11152691]
88. Vu DM, Myers JK, Oas TG, Dyer RB. *Biochemistry.* 2004; 30:3582. [PubMed: 15035628]
89. Uversky VN. *Protein J.* 2009; 28:305. [PubMed: 19768526]

Fig. 1. Experimental FT-IR spectrum of α -synuclein in water at 25 °C. The amide-I band region is shown. The area under the curve was maintained constant in all computed spectra shown in this article.

Fig. 2. Representative clusters of conformations of α -synuclein in terms of backbone solvation (SASA) and secondary structure (percent helicity). A thousand conformations in each of these clusters were chosen to show the dependence of amide I-band on the backbone SASA and helical content. Three-dimensional pictures of representative conformations of α -synuclein for each cluster are shown as well.

Fig. 3. Sensitivity of amide-I band to backbone SASA and helical content of α -synuclein. Computed amide-I band for four distinct conformational clusters (collapsed/low-SASA & coil/low-helical content (green); collapsed/low-SASA & high-helical content (red); extended/high-SASA & coil/low-helical content (cyan); extended/high-SASA & high-helical content (purple) are shown. The measured amide-I band (blue) is shown for comparison. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Frequency shifts from backbone carbonyl oxygen hydrogen bonding that contribute to the amide-I bands of conformational clusters shown in Fig. 3. Total frequency shift (A) due to carbonyl oxygen hydrogen bonding to solvent (B), backbone (C), and side chains (D) are shown. The coloring scheme is explained in Fig. 3.

Fig. 5. Amide-I band for α -synuclein predicted from Bayesian ensemble method is compared to the measured spectrum under different error limits. Predicted spectra along with corresponding Bayesian weights obtained using enthalpically determined priors are shown for low error/ least squares fit (A, B), high error (C,D), and medium error (E, F) limits.

Fig. 6. Accuracy of predicted chemical shifts for backbone atoms based on Bayesian ensemble method with different error limits. Comparison of the chemical shifts predicted for a carbonyl carbon (A, B, C, D), nitrogen (E, F, G, H) and hydrogen (I, J, K, L) atoms of the backbone amide group and backbone C_α atoms (M, N, O, P) are shown for low, medium and high error (enthalpy and entropy) limits.

Table 1

The number of conformations (in units of 100 conformations) belonging to each cluster in all 100 MD simulations of -synuclein.

	Helical content (%)													
	0.0–5.6	5.6–11.2	11.2–16.8	16.8–22.4	22.4–28.0	28–33.6	33.6–39.2	39.2–44.8	44.8–50.4	50.4–56.2				
SASA (nm ²)	72.8–75.6	1.0	0	3.2	7.2	22	23	3.6	1.6	0	0			
	75.6–78.4	1.2	2.0	19	100	167	167	48	2.8	0	0			
	78.4–81.2	90	508	2024	2878	959	224	24	4.1	0	0			
	81.2–84.0	2950	2533	6973	97345	3682	502	21	7.9	9.0	4.0			
	84.0–86.8	12981	3832	4585	5708	2417	446	55	22	7.3	1.3			
	86.8–89.6	16227	1902	990	1092	598	218	33	4.8	3.0	0			
	89.6–92.4	8058	560	124	63	14	2.5	1.1	0.1	0	0			
	92.4–95.2	3966	124	9.2	0.3	0.4	0.1	0.2	0	0	0			
	95.2–98.0	1729	23	1.0	0.2	0.2	0.9	0.3	0	0	0			
	98.0–100.8	122	3.0	0.3	0.1	0.2	0.1	0	0	0	0			

Table 2

Comparison of predicted and experimentally determined chemical shifts. The root mean square deviation (RMSD) and correlation coefficient (Corr Coeff) between the predicted and measured chemical shifts are given. Comparison is carried out for ensembles with Bayesian fits under low, medium, and high (enthalpy and entropy priors) error limits. Chemical shifts are calculated for backbone atoms corresponding to carbonyl C atom, N and HN atoms of the amide group, and the C atoms. SHIFTX is utilized to calculate the chemical shifts for three-dimensional structures in various ensembles and the corresponding accuracy of SHIFTX for the different predictions are given as well. Refer Fig. 6 for further details.

Backbone atoms	Low error			Medium error			High error (enthalpy)			High error (entropy)			SHIFTX accuracy	
	RMSD (ppm)	Corr Coeff		RMSD (ppm)	Corr Coeff		RMSD (ppm)	Corr Coeff		RMSD (ppm)	Corr Coeff		RMSD (ppm)	Corr Coeff
C	1.09	0.80		1.16	1.25	0.64	1.25	0.64		1.25	0.57		1.16	0.89
N	3.83	0.91		0.92	3.61	4.31	0.90	4.31		0.91	4.15		0.91	2.43
HN	0.26	0.48		0.23	0.58	0.29	0.43	0.29		0.41	0.30		0.23	0.58
C	0.73	0.991		0.59	0.994	0.99	0.98	0.99		1.05	0.99		0.98	0.98