Chemical Physics 422 (2013) 143–155

Contents lists available at SciVerse ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

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], S. Gnanakaran [a,*]

[a] Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA
[b] Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA
[c] Physical Chemistry and Applied Spectroscopy, Los Alamos National Laboratory, Los Alamos, NM, USA

ARTICLE INFO

Article history:
Available online 20 May 2013

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

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 μs to characterize the conformational ensemble of α-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.

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 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 α-helices, β-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-depen-

* 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.
E-mail address: gnana@lanl.gov (S. Gnanakaran).

0301-0104/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemphys.2013.05.005

dent 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 D_2O, 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. β-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