

Article

Structural Disorder of Folded Proteins: Isotope-Edited 2D IR Spectroscopy and Markov State Modeling

Carlos R. Baiz¹ and Andrei Tokmakoff^{1,*}¹Department of Chemistry, James Franck Institute and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois

ABSTRACT The conformational heterogeneity of the N-terminal domain of the ribosomal protein L9 (NTL9₁₋₃₉) in its folded state is investigated using isotope-edited two-dimensional infrared spectroscopy. Backbone carbonyls are isotope-labeled (¹³C=¹⁸O) at five selected positions (V3, V9, V9G13, G16, and G24) to provide a set of localized spectroscopic probes of the structure and solvent exposure at these positions. Structural interpretation of the amide I line shapes is enabled by spectral simulations carried out on structures extracted from a recent Markov state model. The V3 label spectrum indicates that the β -sheet contacts between strands I and II are well folded with minimal disorder. The V9 and V9G13 label spectra, which directly probe the hydrogen-bond contacts across the β -turn, show significant disorder, indicating that molecular dynamics simulations tend to overstabilize ideally folded β -turn structures in NTL9₁₋₃₉. In addition, G24-label spectra provide evidence for a partially disordered α -helix backbone that participates in hydrogen bonding with the surrounding water.

INTRODUCTION

Crystallographic protein structures often leave the impression that proteins exist as a unique three-dimensional arrangement of atoms; however, even well-folded proteins are soft, dynamic molecules that sample a range of conformations (1–5). Fluctuations, structural disorder, and conformational changes on timescales from picoseconds to seconds are key elements of protein function, including signaling (6), allostery (7), enzyme catalysis (8), recognition and binding (9,10), and protein-protein interactions (11). Even so, the role of conformational diversity remains underappreciated as rapidly-interconverting structures are still invisible to most experimental techniques. NMR, in addition to its accepted role in protein structure determination, has also provided key insights into protein dynamics (8,12). For example, NMR has revealed the importance of entropic contributions from side-chain fluctuations in defining the relationships among protein folding, stability, and function (9,13,14). Single-molecule methods have provided access to the kinetics of large-amplitude conformational changes on millisecond-to-second timescales (15,16). Recently, ultrafast two-dimensional infrared (2D IR) spectroscopy has emerged as a powerful technique to observe fast conformational dynamics (17–22).

In an effort to experimentally characterize conformational variation in a folded protein, we have made use of amide I 2D IR spectroscopy to investigate the native state ensemble of the 39-residue N-terminal domain of the L9 protein (NTL9₁₋₃₉), shown in Fig. 1 (23,24). The 2D IR

line shapes capture a snapshot of the three-dimensional structure of the backbone ensemble with picosecond resolution, free of motional averaging (17,25). Site-specific structural information is obtained by isotope-labeling of specific backbone carbonyls. A single ¹³C=¹⁸O replacement shifts the frequency of a residue, isolating it from the remaining backbone absorption band. The measured vibrational frequencies of the label, determined by hydrogen bonding to the oxygen, are thereby sensitive to the type and strength of a hydrogen bond as well as fluctuations and solvent exposure. Pairwise labeling is sensitive to the vibrational couplings between two residues (20,26,27). Strong electrostatic interactions between hydrogen-bonded pairs produce large couplings that red-shift (decrease) the vibrational frequency. For NTL9₁₋₃₉, we incorporate ¹³C=¹⁸O isotope labels on the peptide unit following the side chain at five sites in different regions of the protein: V3 (β -sheet), V9 (β -turn), V9G13 (β -turn dual label), G16 (β -strand, solvent-exposed), and G24 (α -helix).

A detailed molecular interpretation of these measurements emerges from structure-based spectroscopic modeling. An atomistic molecular structure is translated into a 2D IR spectrum using a semiclassical model for the coupled amide I dipoles of the backbone that maps the molecular electric field acting on the C=O bond to vibrational frequencies. Dipole-dipole interactions between peptide groups determine the strength of the couplings. Structural variation can be tested by averaging over an ensemble, which can influence linewidth and intensities. As a basis for characterizing structural heterogeneity, we make use of a Markov state model (MSM) that groups conformers from a molecular dynamics (MD) simulation based on exchange kinetics, forming a convenient basis for comparison

Submitted October 8, 2014, and accepted for publication December 19, 2014.

*Correspondence: tokmakoff@uchicago.edu

Editor: H. Jane Dyson.

© 2015 by the Biophysical Society
0006-3495/15/04/1747/11 \$2.00



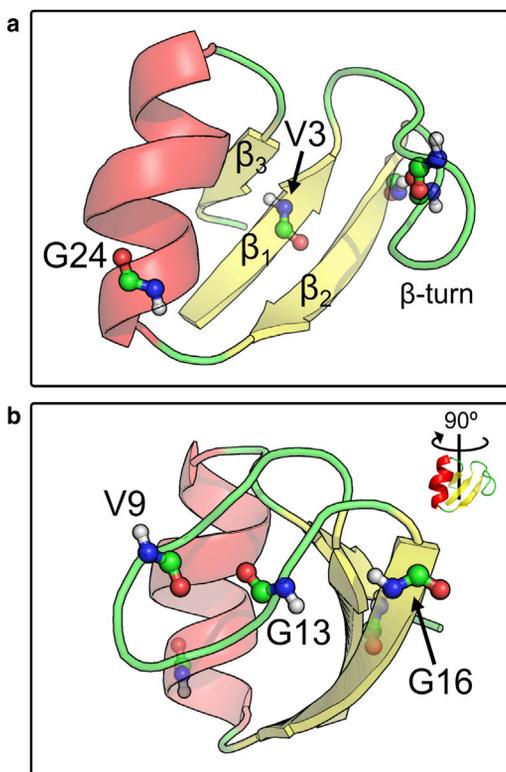


FIGURE 1 (a) Cartoon representation of the crystal structure of NTL9₁₋₃₉ K12M (Protein Data Bank 2HBA). (Red, yellow, and green colors) The α -helix, the β -sheet, and the β -turn regions, respectively. The amide units with $^{13}\text{C}=^{18}\text{O}$ isotope labels are indicated on the structure. (b) Rotated view showing the G16 label as well as the V9 and G13 labels in the β -turn region. To see this figure in color, go online.

between MD simulations and experiments (28–30). The combination of isotope-edited 2D IR spectroscopy, MSMs, and spectral simulations that is used here has recently been used to characterize conformational variation and folding in β -hairpins and β -turns (26,31).

MATERIALS AND METHODS

Isotope labeling and peptide synthesis

Isotope replacement reactions were carried out following a previous protocol (32). Briefly, $^{13}\text{COOH}$ -labeled Fmoc-glycine and valine (Cambridge Isotope Laboratories, Andover, MA) were dissolved in a 6:1 mixture of dioxane: H_2^{18}O , acidified with acetyl chloride and refluxed under a nitrogen atmosphere for ~24 h. Once the reaction had reached completion, the solvent was removed by lyophilization and the product recrystallized from a mixture of dichloromethane and *n*-hexanes. Enrichment ratios of >95% and >99% for Gly and Val, respectively, were confirmed by electrospray ionization mass spectrometry. Peptide samples (sequence: NH₂-MKVIFLKDVKGKGGKGEIKNVADGYANNFLFKQGLAIEA-CONH₂) were synthesized at the Swanson Biotechnology Center (Koch Institute for Integrative Cancer Research at MIT, Massachusetts Institute of Technology, Cambridge, MA). After HPLC purification, samples were triply lyophilized from D₂O (~1 mg/mL) in order to remove residual trifluoroacetic acid and exchange labile protons. Circular dichroism was carried out on the unlabeled sample to confirm proper folding.

Sample preparation and 2D IR spectroscopy

Peptide samples were reconstituted in 10 mM MOPS buffer in D₂O with 100 mM NaCl. The final concentration was kept <5 mg/mL to minimize aggregation. The pH (raw reading, uncorrected for isotope effect) was adjusted with DCl to 2.2 ± 0.1 to protonate all side-chain carboxylates and prevent spectral overlap with the isotope labels by blue-shifting the C=O stretches. Samples were held between two 1-mm CaF₂ windows separated by a 50- μm Teflon spacer in a temperature-controlled cell at $24 \pm 0.1^\circ\text{C}$. Based on the thermodynamic parameters of Horng et al. (24) we estimate a 6–10% unfolded population under these experimental conditions. Because NTL9₁₋₃₉ is prone to aggregation, samples were discarded when aggregation was detected as a sharp IR band at 1612 cm^{-1} , where the detection threshold was ~3% of the molecular concentration. Sample solutions were either prepared fresh or were flash-frozen in liquid nitrogen and stored at -20°C .

The custom-built 2D IR spectrometer was described previously in Khalil et al. (33). In short, spectra were collected in the parallel (ZZZZ) polarization condition. The population time (t_2) was set to 150 fs, and the coherence time (t_1) was scanned to 2.4 and 1.8 ps in 4 fs steps for rephasing and nonrephasing, respectively. Because no off-diagonal features are observed in the isotope-label region, only diagonal slices of the 2D IR spectra are analyzed. Diagonal cuts are offset by $+12\text{ cm}^{-1}$ along the detection axis (ω_3), corresponding approximately to the ridges of maximum intensity in the two-dimensional peaks. Spectra were phased by minimizing differences between the ω_3 projections of the 2D IR spectra and pump-probe spectra in the main amide I band region. The ω_3 offset of the labeled peaks is likely due to a small phase mismatch between the main band and the low-frequency isotope region or slight miscalibration of the detection frequency. Unfortunately, the low signal of the isotope labels in the pump-probe spectrum was not sufficient to reliably phase the spectra using the isotope peaks. To compare the intensities of the different labels, 2D IR spectra are corrected for the spectrum of the pulses and normalized to the maximum intensity of the main amide-I band.

MSM and spectral simulations

The MSM and spectral simulation methods are described in Baiz et al. (34). In summary, the MSM was built from ~2 ms of MD trajectories carried out on the K12M mutant of NTL9₁₋₃₉ by Lindorff-Larsen et al. (35). The full MSM can be generally partitioned into two ensembles: states with low root-mean-square deviation (RMSD) (0–5 Å) to the crystal structure (Protein Data Bank 2HBA) that we call native states, and a set of high-RMSD disordered states (RMSD ~ 5–20 Å) that are separated by a ~2 kcal/mol free-energy barrier (see Fig. 8 in Baiz et al. (34)). Approximately 140 Markov states fall into the native state ensemble, and because the focus of the article is to characterize the heterogeneity of this ensemble, spectral simulations are carried out using the 140 lowest-RMSD states of the full 727-state MSM.

Because the backbone in the native state is heterogeneous, we use the term “folded state” to refer to residues in crystal-structure-like configurations primarily defined by their ϕ/ψ angles and hydrogen-bonding patterns. “Disordered state” refers to local backbone configurations showing solvent-exposed residues, which in the crystal structure appear in stable protein-protein hydrogen bonds. Specifically, 75% of the native population in the MSM has RMSD within 1 Å of the crystal structure. However, it is important to consider that because the ensembles retains most of the native secondary structure, both folded and disordered configurations belong to the native-state ensemble.

Amide I spectra were calculated using a mixed quantum-classical model that maps parameters from an MD simulation to a local mode Hamiltonian that describes the amide I spectroscopy (36). Position-restrained trajectories were launched from five randomly selected configurations of each Markov state (1 ns, 2 fs steps, snapshots saved every 1 ps, SPC/E water, OPLS/AA, using the software GROMACS, Ver. 4.5.1, www.gromacs.org). For each

starting structure, 1000 protein and solvent configurations were converted into local amide I Hamiltonians. Diagonal elements (or site energies) were obtained by mapping the electric field and potential at each residue's N, H, C, O positions into a frequency and transition dipole moment using a map parameterized by Jansen et al. (37,38). Uniform -60 cm^{-1} shifts were applied to site energies corresponding to $^{13}\text{C}=^{18}\text{O}$ labeled units in the Hamiltonian matrices. Off-diagonal terms (couplings) were computed from ϕ/ψ maps for adjacent residues and a transition charge-coupling model for the remaining interactions. Static 2D IR spectra were computed by diagonalization of the Hamiltonian matrices followed by numerical sums over Liouville pathways (39). Spectra were computed as a population-weighted sum over the 140 Markov states. Equilibrium populations were extracted from the first eigenvector of the full MSM as described previously in Lindorff-Larsen et al. (35).

RESULTS

Energy landscape of the native-state ensemble

Fig. 2 shows a network diagram of the 140-state MSM for the native ensemble. States are represented by nodes, which are color-coded by the all-atom RMSD to the crystal structure as well as by the energy of the states extracted from the populations of the full MSM. Node sizes are proportional to the number of connections to other nodes, and the thickness of the connecting lines represents their relative interconversion probabilities. Although some states do not appear to be connected to other states in the network, in the full MSM they are connected to disordered states that are not included in this plot (34). A few low-RMSD states display multiple connections whereas most states have a small number of connections. This is referred to as a “kinetic hub topology”, where low-RMSD states act as hubs for channeling flow through the network (40,41). Each node is, on average, connected to 5.2 other nodes. The top three nodes, which are also the lowest-RMSD states, have degrees of 118, 108, and 71, while high-RMSD nodes are only connected to a few others. From an energy landscape perspective, these kinetic hub states represent a set of lowest energy minima, thus interconversion between states in the ensemble is funneled through these few lowest-RMSD states. The second map, color-coded by pseudo free energy, shows that most of the population resides in low-RMSD states, with a sparse population in the high-RMSD regions. More

specifically, 66% of the total population resides within the four most populated states, with the remaining 34% distributed over 136 states. These structurally similar states represent the mean folded structure of the protein.

These four lowest-RMSD states, labeled A–D, are similar in structure but represent distinct minima in the energy landscape. The four states are within 1.5 kT of each other at the simulation temperature. Structural differences between these states are reflected primarily in the side-chain configurations. For example, in state A, the V9 (β -turn) forms hydrophobic contacts with L30 (α -helix) and F5 (β -sheet), whereas V9 in the state B site is partially solvent-exposed. From the perspective of the backbone, these four configurations display very similar protein-protein hydrogen-bonding patterns, and only small differences in ϕ/ψ angles. Spectroscopically, these low-RMSD states appear virtually identical, thus we refer to these as folded structures. Nonetheless, differences in these structures suggest that the hydrophobic core of NTL 9_{1-39} is markedly heterogeneous and dynamic.

Amide I isotope spectra: experiment

Fig. 3 shows a comparison of experimental and simulated 2D IR spectra of the unlabeled and five different isotope-labeled samples in the $1560\text{--}1620\text{ cm}^{-1}$ region in which the isotope-labeled peaks appear. Above 1600 cm^{-1} the onset of the main amide I band is observed. The anti-diagonal widths are similar for all labels and no cross-peaks are observed between the label and the main band. Therefore, the essential information is captured by the diagonal slices of the 2D IR spectra along the positive peak, as illustrated in Fig. 4. Fourier transform infrared spectroscopy (FTIR) spectra for these samples were uninformative because, except for V3, the low-amplitude label peaks are difficult to distinguish from the strong sloping background in this region.

The spectrum for V3 is typical of what one expects for the $7\text{--}15\text{ cm}^{-1}$ linewidth and Gaussian line shape for a single conformer (26). Remarkably, the other spectra exhibit asymmetric line shapes with widths exceeding 20 cm^{-1} .

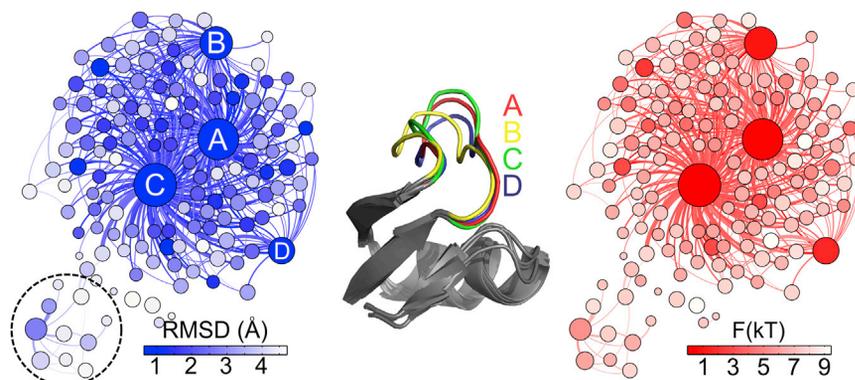


FIGURE 2 (Left) Network plot of the Markov state model for the native ensemble of NTL 9_{1-39} . The states are color-coded by RMSD to the crystal structure. The node size encodes the degree, or number of connections to other states, of each state. (Dashed circle) Registry-shifted states (34). (Center) Structure overlay of the four highest populated states. The main differences are observed in the β -turn region. (Right) Nodes color-coded by the pseudo free energy of native states $F = -\ln(P)$, where P denotes the respective populations in the full MSM. To see this figure in color, go online.

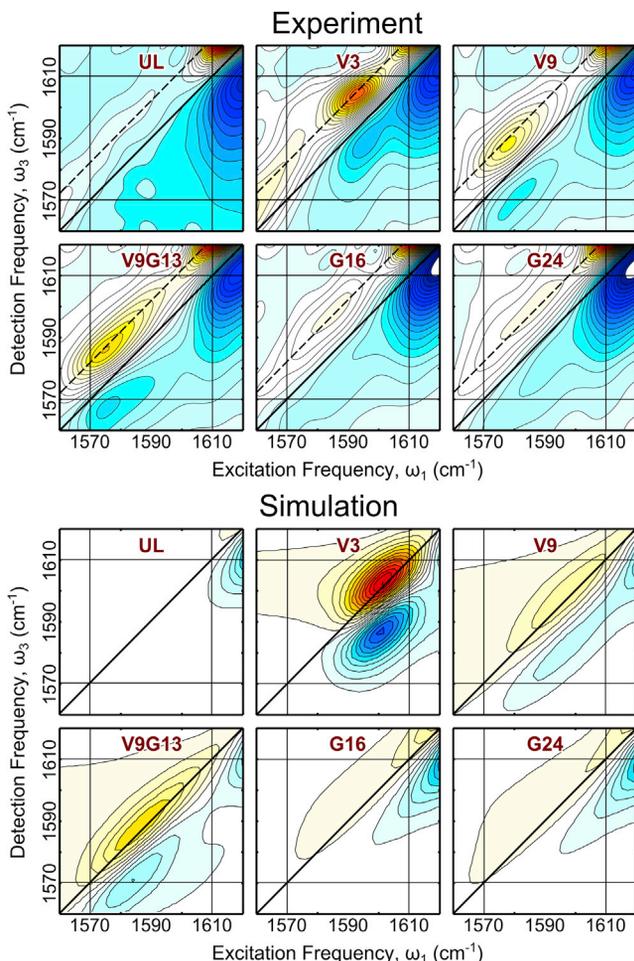


FIGURE 3 Two-dimensional infrared spectra in the label region of the five different isotope labels of NTL9₁₋₃₉. (Dashed lines in the experimental spectrum) Positions of the diagonal slices. To see this figure in color, go online.

For comparison, the average frequency downshift induced by a single hydrogen bond to a backbone carbonyl is $\sim 16 \text{ cm}^{-1}$ (42). Shoulders appear on the red or blue side of the main peaks, indicating the presence of two or more partially overlapping peaks that arise from spectroscopically distinct structural ensembles. Experimental 2D IR diagonals are fit to a sum of Gaussians (Fig. 4, solid curves),

$$\text{fit}(\omega) = \sum_{i=1}^3 A_i e^{-\left[\frac{(\omega - \omega_{c,i})}{\sigma_i}\right]^2},$$

where A_i , $\omega_{c,i}$, and σ_i , represent the amplitude, center frequency, and peak width, respectively (see Table 1). Label spectra are well fit to a sum of two Gaussians with a third component representing the band above 1600 cm^{-1} . Coincidentally, the 12–15 cm^{-1} separation observed between the band centers approximately corresponds to the shift arising from the formation of a hydrogen bond. Also, the linewidth of the lower frequency band broadens

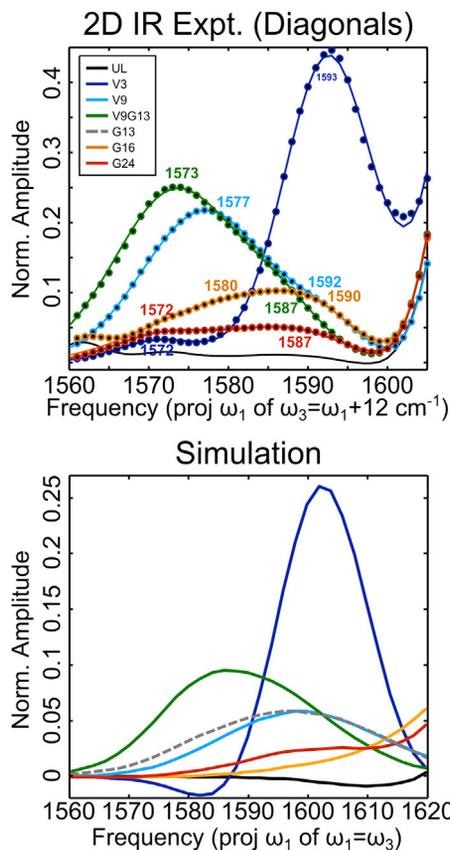


FIGURE 4 Diagonal slices of 2D IR spectra along the peak maxima in the isotope-label region (see Fig. 3). (Top) Experimental spectra. (Circles) Interpolated 2D IR data; (solid curves) dual Gaussian fits to the data as shown in Table 1. Gaussian center frequencies are indicated below each peak. All spectra are normalized to the maximum amplitude of the main amide I band. (Bottom) Simulated diagonal slices of the 2D IR spectra for the same labels. The G13 label (dashed curve) does not appear in the experimental spectrum, but it is included for comparison with V9 and V9G13 experimental curves. To see this figure in color, go online.

relative to the high-frequency band, as expected for a solvent-exposed carbonyl.

The features observed in experimental spectra are as follows.

UL

The unlabeled NLT9₁₋₃₉ spectrum shows no peaks in the label region except for a small feature near 1565 cm^{-1} ,

TABLE 1 Gaussian fit parameters corresponding to the experimental label spectra (Fig. 4)

Parameter	Label				
	V3	V9	V9G13	G16	G24
A_1	3.3E-02	2.2E-01	2.5E-01	8.5E-02	4.0E-02
$\omega_{c,1} (\text{cm}^{-1})$	1572	1577	1573	1580	1572
$\sigma_1 (\text{cm}^{-1})$	8.2	11.1	10.8	13.9	9.0
A_2	4.4E-01	4.8E-02	6.6E-02	4.6E-02	4.9E-02
$\omega_{c,2} (\text{cm}^{-1})$	1593	1592	1587	1590	1587
$\sigma_2 (\text{cm}^{-1})$	7.8	6.7	7.5	7.6	9.9

most likely because of residual deprotonated carboxylate groups ($pK_a \sim 4\text{--}4.5$). The spectrum is included to demonstrate the flat baseline and excellent solvent background suppression of 2D IR. The onset of the main amide I band from the unlabeled vibrations can be observed above 1600 cm^{-1} . Note that the clean baseline observed in the UL spectrum eliminates the need for background subtraction.

V3

The V3 spectrum shows an intense, narrow peak centered around 1593 cm^{-1} along with a low-amplitude peak near 1572 cm^{-1} . The narrow peak indicates that V3(C=O) is held rigidly in place by a single strong hydrogen bond across the β -strands. V3 is located in a largely hydrophobic patch of the protein with minimal solvent exposure.

V9 and V9G13

The V9 spectrum shows two peaks centered around 1577 and 1592 cm^{-1} with 82 and 18% amplitude ratios, respectively. The V9G13 spectrum contains similar features compared to V9 that are red-shifted to 1573 cm^{-1} and 1587 cm^{-1} , but with lower intensity on the blue side and higher intensity on the red side of the band. The increase in intensity is consistent with vibrational coupling between the two residues that results from a pair of hydrogen bonds in the folded β -turn conformation.

G16

The G16 spectrum exhibits broad bands with significantly lower intensity compared to the V3, V9, and V9G13 labels. G16 is located in flexible, highly solvent-exposed regions of the backbone in which G16 carbonyl remains fully solvent-exposed even in the crystal structure (Fig. 1), partially accounting for the observed broader, low-intensity bands.

G24

The G24 spectra display two low-intensity broad peaks, suggesting some heterogeneity in this region. The low intensity of the peak in the simulations (Fig. 4) also point to a disordered and solvent-exposed region.

Amide I isotope spectra: simulations

In comparing experimental and simulated spectra, there are a number of qualitative similarities and some notable differences. For example, the lower intensity of the G16 and G24 labels compared to V3, V9, and V9G13 is successfully captured by the model. Relative intensities of the peaks and the red shift of V9G13 relative to V3 are qualitatively reproduced, but linewidths and peak frequencies are not consistent with experiment. It is important to note that the electrostatic map reproduces the label frequencies and line shapes only qualitatively. Maps provide a means to interpret relative intensities and frequency shifts for

different structures including protein-protein contacts and protein-water hydrogen bonding, but at present cannot provide absolute frequencies or intensities. Aside from this specific shortcoming of the spectral models, discrepancies also arise from structural differences between MSM and the experimental ensembles and thus form the basis for critically evaluating populations in the MSM and experimental ensembles.

UL

Similar to experiment, the unlabeled spectrum shows a smooth baseline with only a small dip in the $1600\text{--}1610\text{ cm}^{-1}$ region.

V3

The V3 label shows a strong, narrow peak, qualitatively similar to experiment. This reflects the single hydrogen bond in the β -sheet, the high backbone rigidity, and stability at this position, and a single spectroscopically distinct state of the backbone. The negative component on the red side of the main band is due to the interference between the positive and negative peaks in the 2D IR spectrum (Fig. 3).

V9, G13, and V9G13

These labels report on the flexibility of the disordered loop region. G13 is included in the simulations to help interpret differences between the V9 and V9G13 experimental spectra. Similar to experiment, V9G13 is red-shifted relative to V9. However, the V9 frequency appears close to that of V3. Interestingly, V9 and G13 display virtually identical peak frequencies and intensities; but V9G13, the dual label, is red-shifted with respect to the individual labels. The shift is attributed to vibrational coupling through a dipole-dipole interaction. Coupling splits the local modes into a red-shifted mode in which the carbonyls oscillate out-of-phase and carry most of the band intensity, and a weak, blue-shifted, in-phase mode. Finally, because the simulated bands are broader compared to experiment, the two peaks are not distinguishable in the simulations. In the next section, we present a more rigorous structural interpretation of the β -turn spectra.

G16 and G24

These two labels show broad, low-amplitude peaks that can be attributed to a high degree of solvent exposure at these residue positions. The separation between high- and low-frequency peaks seen in experimental spectra is not present in the simulations, likely due to a combination of spectral overlap between the broad peaks and, as discussed below, because the MSM may not fully capture the structural ensembles present in the experiment. To characterize the structures that give rise to the distinct peaks in the G24 spectra, we decompose the MSM into different hydrogen-bonding environments near the first turn of the α -helix and compare their spectral signatures.

Backbone heterogeneity

β -turn configuration

The two peaks observed in the experimental V9 and V9G13 spectra alone provide strong evidence for two distinct β -turn ensembles, but spectral modeling based on the MSM delivers more detailed structural insight. MSM configurations reveal two different ensembles: a primarily folded β -turn, accounting for 91% of the MSM population; and a disordered turn, with the remaining 9% population as shown in Fig. 5 *a*. Folded turn configurations show a protein-protein hydrogen bond between the V9(O) and M12(H). In contrast, disordered turn configurations have a solvent-exposed V9(O) acting as a water hydrogen-bond acceptor. The disordered loop states correspond approximately to the near-native state recently described by Schwantes and Pande (43) based on a different MSM derived from the same MD trajectories. According to the new MSM, the population of the near-native state is $\sim 3.5\%$, somewhat lower than our estimated value of 9%. To investigate how these structures are reflected in the spectroscopy, the 140 Markov states were separated into two subensembles based on their hydrogen-bond distance,

$$\vec{r}_{VM} = \left| \vec{r}_{V9(O)} - \vec{r}_{M12(H)} \right|,$$

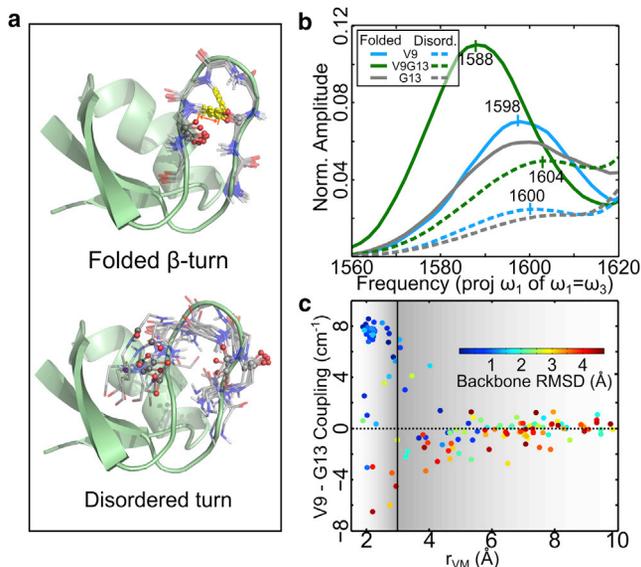


FIGURE 5 (a) Representative structures of the folded β -turn and disordered turn conformations represented in the MSM of NTL9₁₋₃₉ overlaid onto a cartoon structure of the Markov state with lowest RMSD to the crystal structure. (Spheres) Carbonyls corresponding to the V9 and G13 residues. (Yellow dashes) The two folded hydrogen bonds. (Orange double-headed arrow) The V9-M12 hydrogen bond used as the order parameter to distinguish between folded and disordered turn structures. (b) V9-, G13-, and V9G13-label spectra calculated for folded (solid) and disordered (dashed) structures, respectively. (c) Scatter plot of the V9-G13 coupling constant in wavenumbers as a function of r_{VM} distance for the 140 states in the MSM. (Circles are color-coded by the overall RMSD of each state to the crystal structure.) To see this figure in color, go online.

with folded and disordered configurations defined as those with a distance of $r_{VM} < 3.5$ Å and > 3.5 Å, respectively. Note that states A–D in Fig. 2 correspond to folded turn configurations. The choice of order parameter is primarily motivated by the electrostatic contributions from the different atoms to the V9 C=O frequency. Because the M12(H) is the closest nonbonded atom to the C=O unit, r_{VM} is a natural choice for the purpose of spectrally separating the structural ensembles. The 3.5 Å cutoff is selected primarily because it separates the two ensembles based on the V9-G13 coupling strength (Fig. 5 *c*).

Scatter plots in Fig. 6 *a* show the average number of V9(O) protein-protein hydrogen bonds (n_{PP}), including backbone and side-chain contacts, and protein-water hydrogen bonds (n_{PW}) as a function of r_{VM} . Here the geometric hydrogen bond definition ($r_{O\dots O} < 3.5$ Å and $\theta_{OOH} < 35^\circ$) was used for all analyses. On average, the folded turn conformations contain 1.7 protein-protein bonds, and only 0.2 protein-water hydrogen bonds. The folded turn is not a typical β -turn with a single ($n + 2 \rightarrow n$) hydrogen bond, but allows V9(O) to accept two hydrogen bonds from the M12 and G13 amide N-H. Increasing r_{VM} beyond 3.5 Å leads to a decrease of n_{PP} along with an essentially linear increase of n_{PW} . In the disordered conformation, n_{PP} (red) decreases to nearly zero, but n_{PW} (blue) only increases to ~ 1.3 – 1.5 . Therefore, the total number of hydrogen bonds $n_{TOT} = n_{PP} + n_{PW}$ decreases in the disordered configurations. Because the number of hydrogen bonds accepted by the carbonyl is correlated with the frequency of the residue, increasing disorder is expected to blue-shift the labeled band. This data is expressed as V9(O) solvent exposure (n_{PW}/n_{TOT}) in Fig. 6 *b*. Backbone solvent exposure in the β -turn is well

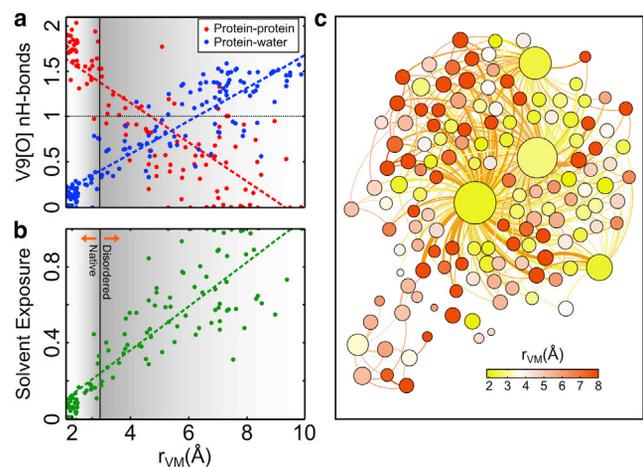


FIGURE 6 Hydrogen-bond analysis of the folded and disordered transitions. (a) Total number of hydrogen bonds accepted by V9(O) separated by protein-protein H-bonds n_{PP} (red) and protein-water H-bonds n_{PW} (blue). (b) Plot of backbone exposure calculated as the ratio of protein-water to total hydrogen bonds $n_{PW}/(n_{PP} + n_{PW})$. (Dashed lines) Linear fits to guide the eye. (c) MSM showing states color-coded by r_{VM} ; (yellow) folded states; (red) disordered states. To see this figure in color, go online.

correlated with r_{VM} , indicating that only folded turns remain protected from the solvent. Finally, in Fig. 6 *c*, the MSM network diagram is color-coded by solvent exposure to show that highly populated states are low-exposure folded turn configurations, similar to the results shown in Fig. 2.

Fig. 5 *b* shows calculated V9, G13, and V9G13 spectra based on folded (*solid curves*) and disordered (*dashed curves*) subensembles. While the widths of the peaks are similar, the folded turns show a $\sim 2\times$ intensity enhancement that results from smaller frequency fluctuations of the residues in protein-protein H-bonds. Analysis of the V9 frequency distributions reveals a standard deviation of 7.0 and 8.5 cm^{-1} for the folded and disordered turns, respectively. In other words, disordered turn V9 exhibits $\sim 20\%$ larger frequency fluctuations due to the high degree of backbone exposure compared to the folded turn. Oscillator strengths for the folded turn V9 are, on average, only 3% higher than the disordered turn V9, revealing that the difference in amplitude results almost exclusively from narrower frequency distributions instead of increases in oscillator strength. Similarly, the folded G13 peaks (*gray curves*) are narrower and more intense compared to disordered turn spectra. The interpretation is the same: folded-turn H-bonds lock the G13 carbonyl into place, lowering solvent exposure. G13 site energy analysis reveals a 12.2 and 15.0 cm^{-1} standard deviation for the folded and disordered configurations, respectively.

V9G13 dual label spectra (*green curves*) are most distinct between the folded and disordered configurations. Without vibrational coupling between the labeled oscillators, V9G13 spectra calculated for disordered configurations are similar to the sum of the individual V9 and G13 spectra, but in the folded turn, vibrational coupling results in a $\sim 10\text{--}12\text{ cm}^{-1}$ red shift and a large intensity increase. The same effect is observed in the experiment where the V9G13 peak appears more intense on the red side and less intense on the blue side compared to V9. The differences between the two labels are, however, smaller in the experiment than simulations. Fig. 5 *c* shows a plot of the V9-to-G13 coupling, namely the off-diagonal term in the Hamiltonian matrix that represents the degree of dipole-dipole coupling between these two sites. The plot indicates that for folded turns, the coupling is $\sim 8\text{ cm}^{-1}$, whereas, for a disordered turn conformation, the coupling is practically zero. This analysis illustrates the short-range nature of amide I vibrational couplings, serving as essentially binary probes of protein contacts.

The calculated V9 band of the folded turn is centered around 1598 cm^{-1} , whereas the disordered turn is centered near 1600 cm^{-1} (Fig. 5 *b*). The small shift can be explained in terms of total hydrogen bonds. To a first approximation, in the native state $n_{TOT} = 2$ and in the disordered state $n_{TOT} = 1.5$, so n_{TOT} decreases by 0.5, resulting in a $\sim 7\text{ cm}^{-1}$ frequency blue shift. The simulations tend to overestimate the spectral shift induced by turn folding (i.e., difference between V9 and V9G13 in simulated

spectra), which could be due to the fact that point dipole interactions, which successfully capture long-range interactions, may overestimate the degree of vibrational coupling between two hydrogen-bonded residues.

Based on these simulations, we believe that the experimental peaks (Fig. 4) centered at 1577 and 1592 cm^{-1} correspond to folded and disordered configurations, respectively. The 15 cm^{-1} difference between these two bands is most likely because simulations underestimate the spectral shifts between the two configurations. Using the intensities from the simulations, we estimate populations of the two states in experiment. Simulated spectra, which are normalized to the main amide I band intensity, show an amplitude ratio of 3.1 between the native and disordered peaks, whereas experiments show a ratio of 4.5. Although we cannot accurately quantify populations with this model, the similarity between these two values indicates that a significant fraction ($>40\%$) of the experimental population is disordered turns, compared to the 9% observed in the MD simulation, indicating that simulations tend to overstabilize folded turn configurations.

Solvent exposure at the G16- and G24-label positions

Next, we examine the solvent exposure at the G16 and G24 positions by binning structures based on the average number of protein-protein and protein-water hydrogen bonds accepted by these carbonyls. For the purpose of analysis, we consider structures that fall into two principal H-bond bins, structures with $n_{PP} \geq 1$ (folded, 89% of total population) and structures with $n_{PW} \geq 1$ (solvent-exposed, 1.4% of total population). Interestingly, the degree of solvent exposure appears to be uncorrelated to the global RMSD of the protein with a correlation coefficient $\rho = 0.06$ (not population-weighted), or to r_{VM} ($\rho = 0.23$) indicating that disorder at the G24 position is purely local in nature. Fig. 7 *a* shows configurations randomly sampled from the two bins, and Fig. 7 *c* shows the MSM network diagram color-coded by G24(O) solvent exposure. The high n_{PP} structures correspond to a folded helix, where the G24(O) is hydrogen bonded to the F29 amide, whereas the high n_{PW} states feature a fully exposed G24 carbonyl. Fig. 7 *b* shows a G24-label spectrum for these two ensembles. The folded spectrum shows a peak centered around 1600 cm^{-1} , whereas the solvent-exposed spectrum only shows the low-frequency edge of the main band. Analysis of the G24 site energies reveals that the average residue frequencies (standard deviations) are 1608 (8.5) and 1610 (7.2) cm^{-1} for solvent-exposed and folded bins, respectively, indicating that the two peaks in the experimental spectrum, 1587 and 1572 cm^{-1} , correspond to a folded helix structure and a solvent-exposed G24 site, respectively. In the spectral models, backbone flexibility contributes to frequency fluctuation though nearest-neighbor frequency shifts, which are calculated via ϕ/ψ angle maps (37,38). In order to partially separate the individual contributions, solvent exposure, and backbone flexibility from the

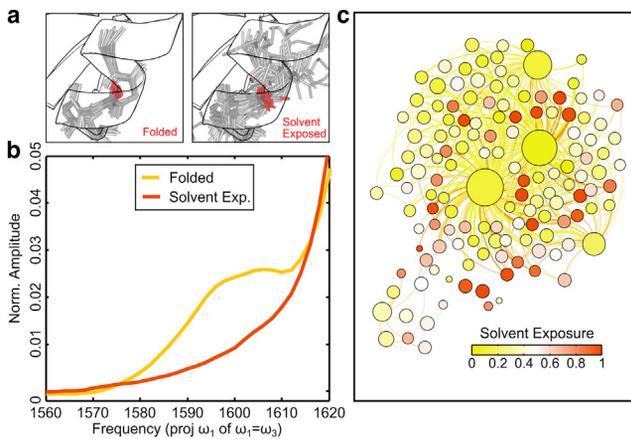


FIGURE 7 (a) Structures of the folded and solvent-exposed G24 configurations of the α -helix aligned to the lowest RMSD structure (*black outline*). (*Red*) The G24 carbonyls. (b) Simulated diagonal 2D IR spectra of the two conformations. The folded peak is more intense compared to solvent-exposed conformations. (c) MSM color-coded by G24(O) solvent exposure: n_{PW}/n_{TOT} . To see this figure in color, go online.

residue frequencies and disordered regions, we repeated the frequency calculations without nearest-neighbor frequency shifts. These frequencies (standard deviations) are 1614 (8.7) and 1599 (7.6) cm^{-1} for solvent-exposed and folded bins, respectively, indicating that the backbone orientation compensates for the blue shift in solvent-exposed configurations. Fluctuations are similar in both sets of simulations, suggesting that the broadness of the line shapes in the G24 spectra originates primarily from solvent exposure.

In the case of G16, which resides in a partially disordered region immediately preceding the third β -strand, we find an average of 1.1 protein-water H-bonds and negligible protein-protein H-bonds, indicating that G16 remains fully solvent-exposed in all structures. These results are not surprising because in the crystal structure, the residue does not display protein-protein hydrogen bonds. Compared with G24, the G16 spectrum has additional intensity on the blue side (Fig. 4), pointing toward a more solvent-exposed backbone at this site.

Finally, it is important to compare the frequency shifts observed for V9 versus G24. In the second label, solvent-exposed configurations are red-shifted compared to the native helix conformation, whereas the opposite trend is observed in the case of V9. These can be interpreted by considering the total number of hydrogen bonds: in the case of G24, the C=O accepts a single hydrogen bond from its partner in the helix, whereas in partially solvent-exposed configurations water molecules participate in additional hydrogen bonds, thus red-shifting the residue frequency. The opposite trend is observed for V9, where folded configurations display two stable protein-protein hydrogen bonds across the β -turn, but only ~ 1.5 protein-water hydrogen bonds in the solvent-exposed structures, explaining the blue shift observed in the spectrum.

DISCUSSION

Structure and stability of folded NTL9₁₋₃₉

The experimental spectra are consistent with a structure that has the overall topology in Fig. 1, and point toward a well-folded β -sheet, a significant population of disordered backbone configurations in the β -turn, and partially disordered residues in the α -helix region. In the experiments, the sharp V3 peak indicates that the residues in β -strands I and II are in rigid hydrogen-bonding configurations with little heterogeneity. In contrast, the breadth and red shift of V9 and V9G13 spectra indicate that a significant fraction ($>40\%$) of the β -turn population resides in disordered and solvent-exposed configurations. The two-peak structure observed for G24 indicates that the first turn of the helix adopts different hydrogen-bonding conformations, perhaps with solvent-exposed carbonyl configurations similar to previous observations for exterior-facing residues of $\alpha_3\text{D}$ (44,45). Spectra for G16 show broad line shapes as a result of disorder and solvent exposure.

Simulations of these spectra using Markov states and other subensembles allow us to build an atomistic picture for the conformations in the folded state that are consistent with experiment. Overall, simulations show a well-folded structure with 66% of the population in four low-RMSD states that are distinguishable only by low-amplitude hinging motions of the β -turn. The remaining 34% of the population has significant configurational freedom, with RMSD values of up to 5 Å. Folded β -sheets characterized by the hydrogen bond at the V3 position are apparent in $>99\%$ of the population, and these populations lead to excellent agreement with experiment. Spectra calculated for other sites indicate that MD simulations underestimate their disordered populations. The MD simulations show a tightly folded β -turn with only $\sim 9\%$ of the population in disordered configurations, whereas our experimental assignment shows a $>40\%$ population of disordered turns. Similarly, solvent-exposed helix turns at G24 constitute 1.4% of the total MSM population, but experiments suggest considerably higher variation in the helix structure. We were unable to assign specific conformations to the G16 line shapes, but the breadth indicates significant disorder and solvent exposure for this carbonyl. Beyond the few lowest-RMSD states, measures of protein disorder in the α -helix and β -loop regions and for the entire protein appear to be uncorrelated, indicating that the disorder arises from local fluctuations with a short correlation length.

Taken together, these results suggest a force field that results in overstabilized structures in the native-state ensemble. However, a one-to-one comparison between simulation and experiment must take into account the following differences in conditions: simulations were carried out on the K12M mutant, a mutation near the center of the loop region that destabilizes the unfolded state with respect to wild-type raising T_m by 18°C. This mutation

could partially account for the overstabilization of the native β -turn configurations in the simulations. Additionally, simulations were run near T_m (for K12M), with multiple folding and unfolding events, yet a coarse separation of the free-energy surface folded and disordered ensembles reveals a folded to unfolded equilibrium constant of ~ 4 (see Baiz et al. (34)), implying that simulations overstabilize the native state ensemble over denatured states. In contrast, experiments were run on wild-type NTL9₁₋₃₉, and at 15°C below T_m (for WT at pH = 2.2) with an equilibrium constant of >10 . Experimental conditions should therefore favor low-RMSD configurations, although the connection between stability (T_m) and backbone flexibility is unclear. Finally, it is important to consider the difference in pH between simulation and experiment. In addition to changes in protein stability, differences in side-chain protonation could alter the electrostatics at the labeled residues and thus shift their frequencies. These pH changes should be taken into consideration in order to get quantitative agreement between simulation and experiment.

These results also add further insight into our previous study of NTL9₁₋₃₉ folding, which was carried out with the full MSM and T-jump protein folding experiments (34). The earlier study, which was rooted in spectral simulations of the fully coupled amide-I vibrations rather than site-specific labels, concluded that the overall changes in calculated 2D IR spectra associated with the global unfolding were consistent with changes observed in experimental spectra. While the earlier work provided evidence for a well-folded secondary structure, these new studies present fresh evidence of backbone solvent exposure that depends on position. Our present interpretation of a stable β -sheet in the MSM appears to be consistent with previous conclusions. Comparing results for the helix suggests that the overall secondary structure is retained in the folded state, but with a significant degree of backbone solvation in the G24 region. This finding points to the importance and power of a full self-consistent analysis of infrared spectroscopy with and without isotope editing strategies.

Further examination of the MSM provides additional predictions regarding the folding mechanism of NTL9₁₋₃₉. A more recent MSM built by Schwantes and Pande (43) from the same MD trajectories showed that 17% of the flux is channeled through near-native states that have a secondary structure similar to the folded state, but which display nonnative side-chain contacts. These states closely resemble the disordered β -turn states described here, suggesting that the turn likely plays a significant role in folding. In addition, the K12M mutation in the β -turn, which is observed to significantly destabilize nonnative states, is observed to participate in the hydrophobic core of the protein. T-jump unfolding experiments with isotope labels in the β -turn would elucidate the relationship between β -turn disorder and the global folding coordinate.

Isotope-edited 2D IR spectroscopy as a structural tool

Our results show that isotope-edited 2D IR spectroscopy and structural modeling based on MD simulations is a promising technique for investigating protein structure in solution. Similar to methods used in magnetic resonance, a self-consistent analysis of multiple structural restraints from site-specific experiments drawing on trial protein structures can be used to determine the most probable structure. However, as a method that captures picosecond snapshots, 2D IR spectra can potentially be extended to characterizing structural heterogeneity, for instance in the study of intrinsically disordered protein structure (31). Our current analysis of experiments is qualitative, focusing on general correlations between structure and amide I frequency shifts and intensities. Further steps will be needed in order to perform quantitative analyses of structural distributions or to build structural ensembles consistent with experiment, most of which are of a theoretical or computational nature.

Of primary importance is the need for quantitatively accurate spectral models. Similar to electronic structure calculations, as of this writing, spectral simulations are useful not as a one-to-one comparison with experiment, but to aid in interpreting the molecular origin of the observed spectra. Spectral models that accurately capture frequency variations in different environments would be essential for quantitative interpretation of experimental spectra. Differences between experimental and simulated spectra of NTL9₁₋₃₉ highlight the challenges associated with applying contemporary electrostatic models to small proteins. Notably, peaks in the simulated spectra appear somewhat broader and blue-shifted compared to experiment. The peak-width differences could be, in part, attributed to motional narrowing, which is not captured by the static models used here. Although the electrostatic map approach has shown excellent agreement with experiment for small peptides (31,46,47), Ganim and Tokmakoff (48) pointed out some general shortcomings that appear to be systematic. Peaks are often blue-shifted with respect to experiment because maps underestimate the solvent-induced frequency shift relative to the gas phase, and the dynamical effects that give rise to line broadening (motional narrowing) are often ignored. Furthermore, systematic studies of vibrational interactions between amide I groups are needed to improve our understanding of the through-bond and through-space coupling mechanisms.

With an accurate spectral model, one can turn toward methods for comparing to trial structures. This will require strategies for identifying statistically meaningful trial structures, which makes the MSM approach particularly appealing. As of this writing, clustering of MSMs relies on structures from MD snapshots, but the fact that IR spectroscopy probes site energies and couplings suggests MSMs could be reencoded into more spectroscopically compatible variables, to more intuitively map structure-frequency

relationships in proteins. Further, methods will be needed to score the information content of an IR spectrum. The frequency and line shape of labeled vibrations or the main amide I band may form strong or limited restraints on a structure. To what extent does a sequence of amide I frequencies restrain a structure, or distinguish between conformers? A statistical approach to this analysis will be required to refine structural ensembles. Strategic selection of specific sites for labeling can be used to design maximally informative experiments.

In principle, the information content that we extracted from our two-dimensional spectra in this study was also available from traditional FTIR spectroscopy. However, the structural interpretation becomes significantly more challenging in practice. The primary advantage is that 2D IR is not affected by solvent absorptions as much as FTIR, and therefore complex difference spectroscopy is not required. This is particularly challenging when conducting specific isotope-labeling studies for large proteins and low concentrations. The same nonlinear effect that suppresses low transition dipole moments in 2D IR spectra also serves to sharpen the resonances of the isotope labels. However, 2D IR has its own unique challenges, including that peak heights are not straightforwardly related to populations, data acquisition times are significantly longer than FTIR, and spectral calculations are significantly longer.

CONCLUSIONS

In summary, isotopic substitutions were used to characterize the hydrogen-bonding environments at different positions along the backbone of NTL9₁₋₃₉ and demonstrate how significant backbone flexibility can be observed for a well-folded protein. The β -loop region in particular, V9 and V9G13, is characterized by a large population of partially disordered states with a fully solvent-exposed backbone. Surprisingly, we observe that the α -helix configurations (G24) still display considerable disorder.

The information content in 2D IR on protein structural heterogeneity and disorder has potential to be used in a variety of contexts, for instance in studies of conformational selection (49,50) or intrinsically disordered proteins (4,51–53). Lacking a well-defined secondary structure, some proteins exist as an ensemble of rapidly interconverting structures that play an active role in biology. Experimentally characterizing protein dynamics has been especially challenging for biophysical methods because crystallography and NMR often reinforce a static, structurecentric view of proteins. A more complete physical description of a protein would not only include the dominant equilibrium conformations but also the free-energy surface around and connecting these minima. Spectra presented here serve to visualize the principal structures that compose the native ensemble, but leave open questions related to kinetics. We expect that, assisted by simulations, T-jump 2D IR experi-

ments on isotope-labeled samples could provide a detailed experimental characterization of the complex free-energy landscape of small proteins (54).

Until recently, points of comparison between simulation and experiment have been at the level of equilibrium structures and folding kinetics. As a result, force fields are built to stabilize secondary structures, and thus may fail to properly sample the configurational space of the protein. Our study illustrates how experimental data can be used to test and improve force fields or aid in developing new experimentally compatible methods for interpreting structural variation seen in MD simulations. In conclusion, we believe that the synergy among state-of-the-art MD simulations, spectral modeling, and ultrafast IR spectroscopy will deliver an intuitive view of proteins beyond average equilibrium structures.

AUTHOR CONTRIBUTIONS

C.R.B. and A.T. designed the research; C.R.B. performed the research; C.R.B. and A.T. contributed analytic tools; C.R.B. analyzed the data; and C.R.B. and A.T. wrote the article.

ACKNOWLEDGMENTS

We thank Christian Schwantes and Professor Vijay Pande (Stanford University, Stanford, CA) as well as Professor Yu-Shan Lin (Tufts University, Medford, MA) for providing the Markov state model and for insightful discussions.

This project was funded by the National Science Foundation (under grants No. CHE-1212557 and No. CHE-1414486), the Massachusetts Institute of Technology Laser Biomedical Research Center (under grant No. P41-EB015871), and a startup grant from the University of Chicago. C.R.B. gratefully acknowledges the National Institutes of Health for a Ruth L. Kirschstein National Research Service Award (under grant No. F32GM105104).

REFERENCES

1. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. *Nature*. 450:964–972.
2. Dill, K. A., and J. L. MacCallum. 2012. The protein-folding problem, 50 years on. *Science*. 338:1042–1046.
3. Bryngelson, J. D., J. N. Onuchic, ..., P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*. 21:167–195.
4. Uversky, V. N., and A. K. Dunker. 2010. Understanding protein non-folding. *Biochim. Biophys. Acta*. 1804:1231–1264.
5. Mobley, D. L., and K. A. Dill. 2009. Binding of small-molecule ligands to proteins: “what you see” is not always “what you get”. *Structure*. 17:489–498.
6. Smock, R. G., and L. M. Gierasch. 2009. Sending signals dynamically. *Science*. 324:198–203.
7. Tzeng, S. R., and C. G. Kalodimos. 2011. Protein dynamics and allostery: an NMR view. *Curr. Opin. Struct. Biol.* 21:62–67.
8. Henzler-Wildman, K. A., M. Lei, ..., D. Kern. 2007. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*. 450:913–916.
9. Boehr, D. D., R. Nussinov, and P. E. Wright. 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5:789–796.

10. Popovych, N., S. R. Tzeng, ..., C. G. Kalodimos. 2009. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proc. Natl. Acad. Sci. USA*. 106:6927–6932.
11. Schreiber, G., G. Haran, and H. X. Zhou. 2009. Fundamental aspects of protein-protein association kinetics. *Chem. Rev.* 109:839–860.
12. Palmer, 3rd, A. G. 2004. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* 104:3623–3640.
13. Wand, A. J. 2001. Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nat. Struct. Biol.* 8:926–931.
14. Baldwin, A. J., and L. E. Kay. 2009. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* 5:808–814.
15. Lu, H. P. 2014. Sizing up single-molecule enzymatic conformational dynamics. *Chem. Soc. Rev.* 43:1118–1143.
16. Schuler, B. 2005. Single-molecule fluorescence spectroscopy of protein folding. *ChemPhysChem*. 6:1206–1220.
17. Ganim, Z., H. S. Chung, ..., A. Tokmakoff. 2008. Amide I two-dimensional infrared spectroscopy of proteins. *Acc. Chem. Res.* 41:432–441.
18. Hamm, P., and M. T. Zanni. 2011. *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge University Press, New York.
19. Serrano, A. L., M. M. Waagele, and F. Gai. 2012. Spectroscopic studies of protein folding: linear and nonlinear methods. *Protein Sci.* 21:157–170.
20. Remorino, A., and R. M. Hochstrasser. 2012. Three-dimensional structures by two-dimensional vibrational spectroscopy. *Acc. Chem. Res.* 45:1896–1905.
21. Kim, Y. S., and R. M. Hochstrasser. 2009. Applications of 2D IR spectroscopy to peptides, proteins, and hydrogen-bond dynamics. *J. Phys. Chem. B*. 113:8231–8251.
22. Fayer, M. D. 2013. *Ultrafast Infrared Vibrational Spectroscopy*. Taylor & Francis, Boca Raton, FL.
23. Horng, J. C., V. Moroz, ..., D. P. Raleigh. 2002. Characterization of large peptide fragments derived from the N-terminal domain of the ribosomal protein L9: definition of the minimum folding motif and characterization of local electrostatic interactions. *Biochemistry*. 41:13360–13369.
24. Horng, J.-C., V. Moroz, and D. P. Raleigh. 2003. Rapid cooperative two-state folding of a miniature α - β protein and design of a thermostable variant. *J. Mol. Biol.* 326:1261–1270.
25. Baiz, C. R., C. S. Peng, ..., A. Tokmakoff. 2012. Coherent two-dimensional infrared spectroscopy: quantitative analysis of protein secondary structure in solution. *Analyst (Lond.)*. 137:1793–1799.
26. Smith, A. W., J. Lessing, ..., J. Knoester. 2010. Melting of a β -hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J. Phys. Chem. B*. 114:10913–10924.
27. Strasfeld, D. B., Y. L. Ling, ..., M. T. Zanni. 2009. Strategies for extracting structural information from 2D IR spectroscopy of amyloid: application to islet amyloid polypeptide. *J. Phys. Chem. B*. 113:15679–15691.
28. Pande, V. S., K. Beauchamp, and G. R. Bowman. 2010. Everything you wanted to know about Markov state models but were afraid to ask. *Methods*. 52:99–105.
29. Noé, F., and S. Fischer. 2008. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* 18:154–162.
30. Voelz, V. A., G. R. Bowman, ..., V. S. Pande. 2010. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J. Am. Chem. Soc.* 132:1526–1528.
31. Lessing, J., S. Roy, ..., A. Tokmakoff. 2012. Identifying residual structure in intrinsically disordered systems: a 2D IR spectroscopic study of the GVGXPGVG peptide. *J. Am. Chem. Soc.* 134:5032–5035.
32. Marecek, J., B. Song, ..., D. P. Raleigh. 2007. A simple and economical method for the production of ^{13}C , ^{18}O -labeled Fmoc-amino acids with high levels of enrichment: applications to isotope-edited IR studies of proteins. *Org. Lett.* 9:4935–4937.
33. Khalil, M., N. Demirdoven, and A. Tokmakoff. 2003. Coherent 2D IR spectroscopy: molecular structure and dynamics in solution. *J. Phys. Chem. A*. 107:5258–5279.
34. Baiz, C. R., Y. S. Lin, ..., A. Tokmakoff. 2014. A molecular interpretation of 2D IR protein folding experiments with Markov state models. *Biophys. J.* 106:1359–1370.
35. Lindorff-Larsen, K., S. Piana, ..., D. E. Shaw. 2011. How fast-folding proteins fold. *Science*. 334:517–520.
36. la Cour Jansen, T., A. G. Dijkstra, ..., J. Knoester. 2006. Modeling the amide I bands of small peptides. *J. Chem. Phys.* 125:44312.
37. la Cour Jansen, T., and J. Knoester. 2006. A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J. Chem. Phys.* 124:044502.
38. la Cour Jansen, T., and J. Knoester. 2006. Nonadiabatic effects in the two-dimensional infrared spectra of peptides: application to alanine dipeptide. *J. Phys. Chem. B*. 110:22910–22916.
39. Cheatum, C. M., A. Tokmakoff, and J. Knoester. 2004. Signatures of β -sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J. Chem. Phys.* 120:8201–8215.
40. Bowman, G. R., and V. S. Pande. 2010. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. USA*. 107:10890–10895.
41. Dickson, A., and C. L. Brooks, 3rd. 2013. Native states of fast-folding proteins are kinetic traps. *J. Am. Chem. Soc.* 135:4729–4734.
42. Gnanakaran, S., and R. M. Hochstrasser. 2001. Conformational preferences and vibrational frequency distributions of short peptides in relation to multidimensional infrared spectroscopy. *J. Am. Chem. Soc.* 123:12886–12898.
43. Schwantes, C. R., and V. S. Pande. 2013. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* 9:2000–2009.
44. Manas, E. S., Z. Getahun, ..., J. M. Vanderkooi. 2000. Infrared spectra of amide groups in α -helical proteins: evidence for hydrogen bonding between helices and water. *J. Am. Chem. Soc.* 122:9883–9890.
45. Walsh, S. T., R. P. Cheng, ..., W. F. DeGrado. 2003. The hydration of amides in helices; a comprehensive picture from molecular dynamics, IR, and NMR. *Protein Sci.* 12:520–531.
46. Woys, A. M., A. M. Almeida, ..., M. T. Zanni. 2012. Parallel β -sheet vibrational couplings revealed by 2D IR spectroscopy of an isotopically labeled macrocycle: quantitative benchmark for the interpretation of amyloid and protein infrared spectra. *J. Am. Chem. Soc.* 134:19118–19128.
47. Reppert, M., and A. Tokmakoff. 2013. Electrostatic frequency shifts in amide I vibrational spectra: direct parameterization against experiment. *J. Chem. Phys.* 138:134116.
48. Ganim, Z., and A. Tokmakoff. 2006. Spectral signatures of heterogeneous protein ensembles revealed by MD Simulations of 2DIR spectra. *Biophys. J.* 91:2636–2646.
49. Kiefhaber, T., A. Bachmann, and K. S. Jensen. 2012. Dynamics and mechanisms of coupled protein folding and binding reactions. *Curr. Opin. Struct. Biol.* 22:21–29.
50. Daniels, K. G., N. K. Tonthat, ..., T. G. Oas. 2014. Ligand concentration regulates the pathways of coupled protein folding and binding. *J. Am. Chem. Soc.* 136:822–825.
51. Wright, P. E., and H. J. Dyson. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293:321–331.
52. Xie, H., S. Vucetic, ..., Z. Obradovic. 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* 6:1882–1898.
53. Habchi, J., P. Tompa, ..., V. N. Uversky. 2014. Introducing protein intrinsic disorder. *Chem. Rev.* 114:6561–6588.
54. Jones, K. C., C. S. Peng, and A. Tokmakoff. 2013. Folding of a heterogeneous β -hairpin peptide from temperature-jump 2D IR spectroscopy. *Proc. Natl. Acad. Sci. USA*. 110:2828–2833.