A Molecular Interpretation of 2D IR Protein Folding Experiments with Markov State Models

Carlos R. Baiz,[¶] Yu-Shan Lin,[†] Chunte Sam Peng,[¶] Kyle A. Beauchamp,[‡] Vincent A. Voelz,[⊥] Vijay S. Pande,[†‡§] and Andrei Tokmakoff[¶*]

[†]Department of Chemistry, [‡]Biophysics Program, and [§]Department of Structural Biology, Stanford University, Stanford, California; [¶]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts; and [⊥]Department of Chemistry, Temple University, Philadelphia, Pennsylvania

ABSTRACT The folding mechanism of the N-terminal domain of ribosomal protein L9 ($NTL9_{1-39}$) is studied using temperature-jump (T-jump) amide I′ two-dimensional infrared (2D IR) spectroscopy in combination with spectral simulations based on a Markov state model (MSM) built from millisecond-long molecular dynamics trajectories. The results provide evidence for a compact well-structured folded state and a heterogeneous fast-exchanging denatured state ensemble exhibiting residual secondary structure. The folding rate of 26.4 μs^{-1} (at 80°C), extracted from the T-jump response of $NTL9_{1-39}$, compares favorably with the 18 μs^{-1} obtained from the MSM. Structural decomposition of the MSM and analysis along the folding coordinate indicates that helix-formation nucleates the global folding. Simulated difference spectra, corresponding to the global folding transition of the MSM, are in qualitative agreement with measured T-jump 2D IR spectra. The experiments demonstrate the use of T-jump 2D IR spectroscopy as a valuable tool for studying protein folding, with direct connections to simulations. The results suggest that in addition to predicting the correct native structure and folding time constant, molecular dynamics simulations carried out with modern force fields provide an accurate description of folding mechanisms in small proteins.

INTRODUCTION

Understanding the molecular interactions that lead proteins to fold into well-defined three-dimensional structures remains extremely challenging, primarily due to a lack of experimental techniques that provide the structural sensitivity and time resolution required to monitor microsecond protein dynamics and their role in folding (1–8). Descriptions of folding are developed under the light of experimental observations, and thus influenced by the methods used to study them (9,1,11,12). Because most experiments generally average over a large number of conformational degrees of freedom, they remain insensitive to the microscopic dynamics of folding. For example, over the last two decades it has been established that observed two-state folding is not an indicator of simple dynamics, but most likely arises from conformationally diverse folding pathways (13–15). Specifically, a configurationally varied unfolded ensemble that samples a network of pathways funneling the system toward a structurally well-defined folded state can result in two-state thermodynamics and kinetics (16–19). Experimental evidence for this picture has emerged from studies of fast folding proteins and downhill folding (16,20).

In parallel, over the last few years, the extension of protein folding simulations into the millisecond regime has provided extensive sampling for a variety of fast folding proteins, and has enabled the construction of new kinetic models that provide a structural view of the folding process on timescales relevant to experiments (21–25). These developments further highlight the need for experimental probes that can directly access molecular aspects of folding over multiple timescales. In brief, connections between theory and experiments have remained weak. In light of these recent theoretical and computational developments, we report on advances that enhance the structural and temporal sensitivity in temperature-jump (T-jump) protein folding experiments by building detailed quantitative comparisons between two-dimensional infrared spectroscopy (2D IR) and molecular dynamics (MD) simulations.

2D IR spectroscopy of backbone amide I′ vibrations is rapidly developing as a powerful technique for characterizing protein structure in solution (3,26,27). Vibrational spectroscopy offers the advantage of sensitivity to global protein conformation (3,26–29), because long-range electrostatic interactions produce delocalized vibrations spanning large portions of the backbone. 2D IR provides enhanced structural resolution over conventional IR absorption spectroscopy by spreading the spectral information onto two frequency axes, excitation (ω_1) and detection (ω_3), to map the couplings between protein vibrations (30–34). Compared to the timescales of protein motion, the subpicosecond time-resolution afforded by the ultrafast laser pulses makes a 2D IR spectrum an instantaneous snapshot of the structural ensemble. These features present opportunities for 2D IR spectroscopy to investigate rapidly interconverting protein ensembles (2). With the

Submitted October 9, 2013, and accepted for publication February 3, 2014.

*Correspondence: tokmakoff@uchicago.edu

Carlos R. Baiz and Andrei Tokmakoff's present address is University of Chicago, 929 E 57th St., Chicago, IL, 60637.

Yu-Shan Lin's present address is Department of Chemistry, Tufts University, Medford, MA 02155.

Editor: Daniel Raleigh.

0006-3495/14/03/1359/12 $2.00

http://dx.doi.org/10.1016/j.bpj.2014.02.008

development of new spectroscopic models, amide I′ IR spectra now can be semiquantitatively simulated from protein structures, providing an avenue to connect with atomistic MD simulations (29,36,37).

MD simulations provide exquisitely detailed structural and dynamical information. Recent developments in computer hardware combined with new algorithms have enabled millisecond-long MD simulations of small proteins, showing multiple folding events in a single trajectory. These simulations reveal the vast number of configurations available to a protein and the complexity of the energy landscape that make modeling of folding processes extremely challenging (25,38,39). Markov state models (MSMs) of these simulations reduce the complexity of the process by discretizing the configurational landscape, and bridge the gap between the atomistic and statistical folding models by representing it as a network of structures—or Markov states—interconnected by their respective interconversion rates (40,41). In essence, the MSM approach is analogous to the kinetic description of coupled chemical reactions most familiar to biochemists, except the models often include thousands of states. In addition to providing an intuitive description of protein folding, MSMs reduce the simulation trajectories to statistical ensembles of molecular structures. Thus, MSMs also provide advantages for analyzing protein conformers present in experiments. To this point, MSMs have remained difficult to validate experimentally as it is challenging to make direct comparisons with experimental measurements beyond equilibrium structures and folding kinetics (42).

This work describes the folding mechanism of the 39-residue fast-folding mixed α/β N-terminal domain of the L9 protein, denoted NTL9$_{1-39}$ (PDB: 2HBA, Fig. 1 *a*). We combine 2D IR spectroscopy with a laser-induced T-jump to probe nonequilibrium protein dynamics—denaturation and refolding—on timescales from microseconds to milliseconds (43). The transient 2D IR spectra that encode conformational changes are compared with spectra simulated from an MSM (44) to build an intuitive structural interpretation of the protein folding process. This approach directly connects experiments and simulations, enabling the validation of MSMs, and providing a new platform to validate force field parameters, particularly in relation to structural heterogeneity and folding (45,46). In turn, simulations allow for direct structural interpretation of IR measurements and contribute to the molecular understanding of protein folding.

MATERIALS AND METHODS

Detailed descriptions of the experimental methods, Markov state models, and spectral simulations have been provided previously or are provided in the Supporting Material. This section summarizes the important information.

Samples

Samples of NTL9$_{1-39}$ were synthesized at the Koch Institute Biopolymers Facility at MIT using FMOC-based solid-state peptide synthesis. Samples were purified by reverse-phase HPLC and their identity and purity was confirmed by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. Protein samples were hydrogen/deuterium exchanged and lyophilized in HCl/D_2O and reconstituted in a 10 mM MOPS, 100 mM NaCl D_2O buffer at pH* = 5.6 (*pH meter reading in D_2O, not corrected for isotope effect). Protein concentrations were kept below 10 mg/mL to minimize aggregation. Circular dichroism and IR absorption measurements confirmed that the protein was properly folded at low temperature. For all IR experiments, the sample was held between two CaF_2 windows separated by a 50 μm Teflon spacer, and mounted into a temperature-regulated brass jacket.

Spectroscopy

Three IR measurements are described in this work: Fourier-transform infrared (FTIR) spectroscopy, 2D IR spectroscopy, and heterodyne-detected vibrational echo (HDVE) spectroscopy. FTIR is a traditional IR absorption measurement and is carried out using a commercial spectrometer. The other measurements are carried out using a homebuilt ultrafast optical setup.

For the purpose of interpretation, 2D IR represents a map of changes in the IR spectrum (detection axis, ω_3) in response to excitation at a given frequency (excitation axis, ω_1). Analogous to 2D NMR spectroscopy, cross-peaks are observed when two vibrations are coupled. In 2D IR, each peak appears as a positive/negative doublet (see Fig. 2, for example), which can be understood as follows: after excitation from the ground to the first excited state ($0 \rightarrow 1$), the detection pulses can either stimulate emission down to the ground state ($1 \rightarrow 0$), which appears as a positive diagonal

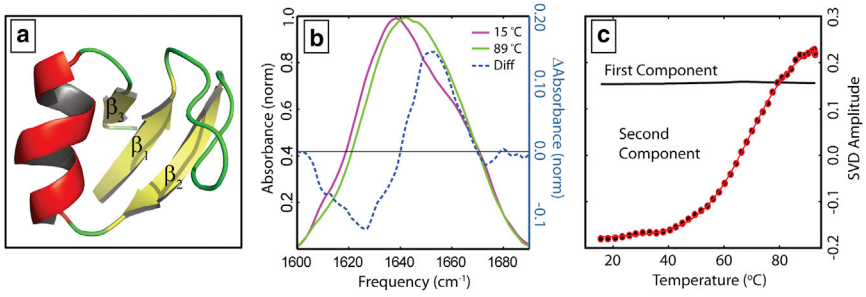

FIGURE 1 (*a*) Crystal structure of NTL9$_{1-39}$ (PDB: 2HBA), (*b*) Temperature-dependence of the amide I′ band of NTL9$_{1-39}$. The magenta and green curves represent the equilibrium absorption at 15 and 89°C, respectively, where NTL9$_{1-39}$ is folded and partially unfolded, respectively. The dashed curve is the difference between the two temperatures (*right axis*). The difference curve shows loss in the low- and high-frequency regions along with a gain in the center region of the spectrum. (*c*) First and second singular value components from the temperature-dependent FTIR. The solid curve represents a two-state fit to the data. Thermodynamic parameters extracted from the fit are described in the text. To see this figure in color, go online.

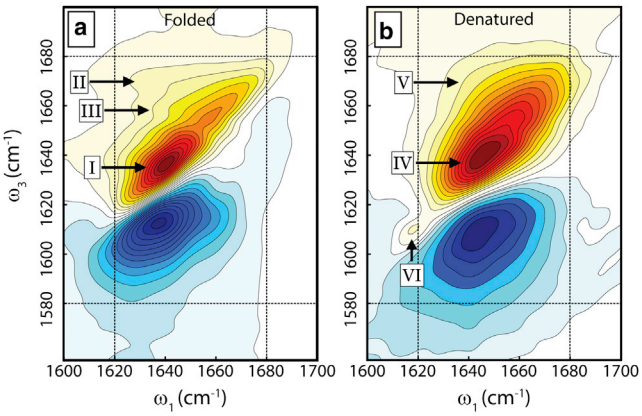

FIGURE 2 Absorptive amide I′ 2D IR spectra at T = 3°C (*a*) and T = 80°C (*b*). Spectra were collected at 150 fs waiting time in the ZZYY polarization condition. Positive contours are indicated in red and negative in blue. Contours are spaced using a hyperbolic sine function to emphasize the low-amplitude features. To see this figure in color, go online.

peak, or induce further absorption to the second excited state (1 → 2), which appears as a negative peak below the diagonal.

Detailed descriptions of the T-jump 2D IR spectrometer can be found in previous publications (47–49). In brief, the output of a 1 kHz Ti:Sapphire laser (792 nm) is used to generate 90-fs pulses centered at 6 μm through optical parametric amplification/difference frequency generation. Pulses are split into four copies with wavevectors k_1, k_2, k_3, and k_{LO}, and focused to a 100 μm spot at the sample. The signal is emitted in a background-free direction. The full electric field of the signal is characterized through spectral interferometry with a reference pulse (k_{LO}). The reference beam is passed through the focus such that the pulse experiences the same changes in index of refraction caused by the T-jump as the excitation and signal fields. 2D spectra are recorded by scanning the delay between the first and second pulses, Fourier-transformation with respect to the first time delay generates the excitation frequency. The detection frequency is measured directly in a grating spectrometer equipped with a 2 × 64-pixel mercury-cadmium-telluride array. All 2D IR and HDVE spectra were recorded with a waiting-time (τ_2) of 150 fs.

An HDVE spectrum has a single frequency axis but retains some of the coupling information of 2D IR. Specifically, the HDVE spectrum is equivalent to the projection of its corresponding 2D IR spectrum onto the detection axis. Both spectra are collected using the same sequence of pulses, but HDVE spectra do not require scanning the time delays between excitation pulses (τ_1), reducing the time required for data acquisition by ~100×. However, because HDVE only reveals a projection of the 2D IR surface onto the detection axis, weak spectral features, such as crosspeaks, can be eclipsed by the intense diagonal peaks, making the interpretation more difficult. Here, we use HDVE spectroscopy to measure the protein kinetics by finely sampling the time response following a T-jump. We measure 2D IR spectra at a few specific delays to provide more detailed structural insight into the evolution of the protein ensemble.

Transient 2D IR measurements are made by measuring changes of the 2D IR spectrum following the T-jump. T-jump pulses were generated by a BBO-based optical parametric oscillator (OPO), pumped by the output of a 20 Hz Nd:YAG laser, producing 11-mJ 7-ns pulses resonant with the overtone of the O-D stretch of the D_2O solvent ($\lambda = 1.98$ μm). By focusing to a 500 μm spot, a spatially uniform T-jump of 10°C is created at the focus of the 6 μm beams. The T-jump laser is electronically synchronized to the 2D IR spectrometer, and the time delay between the T-jump and the 2D IR pulses can be electronically scanned from nanoseconds to milliseconds. Spectral changes induced by the T-jump are plotted as difference spectra before and after the T-jump. The response of the solvent is simultaneously measured by the changes in transmission of the reference pulse through the sample. After the T-jump, heat diffusion out of the sample into the CaF_2 windows causes the sample to relax to its original temperature with a ~2.7 ms decay. The longest delay we measured (49 ms) is given by the repetition rate of the T-jump laser (20 Hz). T-jump spectra were collected in the all-parallel (ZZZZ) polarization geometry at two initial temperatures, 60 and 70°C, around the melting point of $NTL9_{1–39}$ (64°C). The excitation frequency was scanned in 14 fs steps (undersampled) from 0 to <2.0 ps and <1.0 ps for rephasing and nonrephasing spectra, respectively.

MSMs

MSM were built from two ~1 ms long MD trajectories of the K12M mutant of $NTL9_{1–39}$ (21,24,44,50), which were provided by Shaw et al. (21) with configurations stored every 50 ns. The trajectories were subsampled every 10 steps and the resulting configurations were clustered using the k-center clustering algorithm (51) based on a root mean-square deviation (RMSD) of heavy atoms. The clustering procedure was performed until the newly identified cluster center had a RMSD <2 Å from the previously identified representatives. A total of 727 cluster centers were identified. Using these cluster centers, state indices were assigned to the full trajectories and transitions between states were counted at a lag time of 750 ns to construct a transition matrix. We point out that this new MSM offers a significant improvement over the previous $NTL9_{1–39}$ MSM (52), primarily due to the long explicit-solvent simulations and the multiple folding-unfolding events observed in the trajectories. We also point out that a recent MSM for $NTL9_{1–39}$ includes dynamical relaxation in the construction of state space and should constitute an improvement over the present MSM (22).

The exchange kinetics between Markov states are represented by the magnitude of each eigenvector component of the Markov transition matrix (41). The first eigenvector describes the time-invariant equilibrium populations, and the second eigenvector represents the changes in populations associated with the slowest structural interconversion in the system, namely the global unfolding. The state with the largest population (state 717) has a C_α-RMSD of ~0.7 Å to the crystal structure (PDB: 2HBA). Sorting the states based on their RMSD to experiment shows that states with large equilibrium population tend to have small RMSD to the experimental native structure. A detailed characterization of the results is presented in the Supporting Material.

Network graphs were used to visualize the MSM. These were generated using the ForceAtlas algorithm as implemented in Gephi 0.81 (53). In brief, this graph represents the MSM as a set of masses (nodes) connected by springs (edges). Each node represents a Markov state. Nodes are connected to other nodes by springs whose spring constant is proportional to the transition rate between those states. Repulsion terms are added to prevent node overlap and the layout is optimized by minimizing the overall energy of the system. States with rapid interconversion appear closer to one another in the optimized layout than states that interconvert slowly.

Spectral simulations

Within the general simulation approach, a spectrum corresponding to a three-dimensional structure can be obtained using a mixed quantum-classical model described in detail elsewhere (54,55). Briefly, a quantum mechanical Hamiltonian for the coupled backbone amide I′ vibrations of the individual peptide linkages (sites) is parameterized using mapping variables that depend on the classical protein structure. Site energies are generated using a frequency map that correlates the electrostatic potential and gradient at the C, O, N, H positions to a vibrational frequency and transition dipole moment (55). Couplings between adjacent bonded sites are generated using nearest-neighbor ϕ/ψ maps, and electrostatic transition charge coupling is used for through space interactions (54). One- and two-quanta energy levels and dipoles are extracted through diagonalization of the

amide I′ Hamiltonian. 2D spectra are generated by transition-dipole-weighted sums over Liouville pathways (28).

To properly account for the solvent electrostatics for a given Markov state, five randomly selected structures were solvated, and a 100-ps-long position-restrained MD simulation (NPT 300 K, SPC/E water, 2 fs steps, snapshots every 1 ps) was run using the all-atom OPLS force field implemented in GROMACS 4.5.1 (56). Calculated spectra were obtained as an average of these configurations within the static ensemble. Final difference spectra are generated by weighting each Markov state spectrum by its corresponding second eigenvector amplitude (see Fig. 5).

RESULTS

Equilibrium characterization of NTL9$_{1-39}$

Amide I′ IR absorption spectra of NTL9$_{1-39}$ at low and high temperatures, namely folded and denatured ensembles, are shown in Fig. 1 *b*. The folded spectrum shows a two-band shape characteristic of α/β proteins, whereas the denatured spectrum shows a single broad band centered around 1640 cm^{-1}. Spectral differences become more evident in the difference plot (dashed curve), which shows negative peaks around 1630 and 1670 cm^{-1}, and a broad gain feature around 1650–1660 cm^{-1}. Therefore, the negative peaks involve the loss of β-sheet configurations in the ensemble, the low- and high-frequency peaks arise from $\nu_\perp$ and $\nu_\parallel$ transitions of the β-sheet, respectively, and the center band is attributed to disordered states. Spectra are consistent with the loss of structure upon thermal denaturation generally observed in globular proteins.

Singular value decomposition projects out the principal spectral components along with their temperature profiles. The first singular value decomposition component represents an average spectrum and remains invariant with temperature (Fig. 1 *c*). Raleigh and co-workers (44,57) have provided extensive thermodynamic and kinetic evidence for cooperative two-state folding in NTL9$_{1-39}$. Here, we fit the thermal denaturation curve with a two-state model to confirm that the sample behaves properly under our experimental conditions. The sigmoidal shape of the second component along with the small amplitude and no clear temperature trend of higher components are consistent with the observed two-state behavior of NTL9$_{1-39}$. The stability curve is fit to a standard thermodynamic function of the form (58):

$$\Delta G = \Delta H(T_m) - T\Delta S(T_m) + \Delta C_P[T - T_m - T(ln(T/T_m))],$$

producing the following values:

$$\Delta H(T_m) = 25.21 \pm 0.59 \text{ kcal mol}^{-1},$$
$$\Delta S(T_m) = 73.6 \pm 1.7 \text{ cal mol}^{-1}\text{ K}^{-1},$$
$$\Delta C_P = 145 \pm 6 \text{ cal mol}^{-1}\, K^{-1},\ T_m = 64 \pm 3^\circ\text{C}.$$

These values compare favorably with published results of Raleigh and co-workers based on circular dichroism denaturation curves (44).

Amide I′ 2D IR spectra of folded and denatured NTL9$_{1-39}$ (Fig. 2) show complex lineshapes with regions of positive and negative amplitude that arise from the overlap of resonances from multiple delocalized vibrations. The positive peak (*red*) arises from the ground to first excited state transitions (0—1) and the negative band (blue) represents transitions between first and second vibrational states (1—2). The low-temperature 2D IR spectrum of NTL9$_{1-39}$ (Fig. 2 *a*) exhibits the features associated with an α/β protein: The intense diagonal peak centered around 1630–1640 cm^{-1} (labeled **I**) and the off-diagonal ridge observed around 1680 cm^{-1} (**II**) are attributed to the $\nu_\perp$ and $\nu_\parallel$ transitions of the β-sheet, a second ridge (**III**) is also observed around 1660 cm^{-1} (59,60). The off-diagonal ridges (crosspeak) arise from the coupling between the two primary β-sheet modes, a characteristic signature of properly folded β-sheets. A diagonal band centered around 1650–1660 cm^{-1} is typically assigned to the nearly degenerate ν_A and ν_{E1} transitions of the α-helix. Thermally denatured spectra (Fig. 2 *b*) show broad diagonal peaks centered around 1640–1650 cm^{-1} (**IV**) with weak off-diagonal features, typically associated with unstructured configurations (3). However, the residual plateau centered around $[\omega_1, \omega_3] = (1640, 1670)$ cm^{-1} (**V**) suggests that a small amount of β-sheet structure remains present in the ensemble at 80°C. The sharp peak at 1612 cm^{-1} (**VI**) is attributed to a small amount of aggregate (see below).

2D spectroscopy has the capability of measuring residual secondary structure and heterogeneity in disordered systems (37,61,62). Singular-value decomposition of the high-temperature spectrum using a benchmark set of proteins described in (31) predicts ~10% of residues in β-sheet conformations in the unfolded state (see Fig. S4). The current measurements cannot distinguish between native populations that remain folded at high temperatures or denatured/misfolded proteins with partial β-sheet character. However, the thermodynamics extracted from the FTIR melting curve (Fig. 1) predict 9% of native population at 90°C. Considering that native NTL9$_{1-39}$ contains 28% β-sheet, we estimate that native β-sheet configurations represent ~2–3% of the total residues (native + denatured). Therefore, the measured 10% residual sheet configuration suggests that ~7–8% of the residues in the denatured state ensemble (DSE) assume β-sheet configurations. These findings are consistent with previous experiments that have provided evidence for a compact DSE (63,64). Secondary structure analysis of our MSM using the DSSP algorithm (65) shows that 6% of residues in the DSE assume β-sheet configurations (21). The close correspondence further validates the simulation and analysis methods, because structural assignment of MD snapshots relies on local backbone φ/ψ angles, whereas IR spectroscopy probes the delocalized vibrational modes that result from long-range order of the residues (66).

T-jump HDVE kinetic measurements

Fig. 3 shows the mean HDVE response in the 1620–1654 cm^{-1} region from 1 μs to 50 ms. The protein shows an initial ~100 μs response with a reequilibration on the ms timescale. The measured curves represent a convolution of the protein unfolding and folding kinetics and solvent temperature reequilibration. The deconvolved data is fit to a single exponential, giving a time constant of 204 $\pm$ 9 μs at $T_i = 60°C$ and 112 $\pm$ 9 μs at $T_i = 70°C$ (Table 1, Fig. S6). Single exponential relaxation is a characteristic kinetic behavior of two-state folders, indicating that the barrier to unfolding is large with respect to k_BT, and barrier crossing rates are significantly slower than the intrawell diffusion.

T-jump 2D IR spectroscopy

Difference 2D IR spectra, shown in Fig. 4, were collected at three different T-jump delays: 1 μs, 100 μs, and 1 ms. Additional spectra were obtained at millisecond intervals up to 48 ms, but aside from an overall amplitude decay, spectral changes are not observed after 1 ms. A fully reequilibrated spectrum at 60°C is plotted at a delay of 49 ms.

According to the T-jump HDVE kinetics (Fig. 3) no structural perturbation is expected at 1 μs. However, difference spectra show broad diagonal bleach with two main features. The first is one intense bleach at low-frequency labeled **a** and **a′** (each corresponding anharmonic peak is labeled with a prime) attributed primarily to a rapid weakening of protein hydrogen bonds that occurs as fast as the temperature is raised. The second is a higher frequency diagonally stretched bleach (**b**) that can be attributed to small shifts in intensity and lineshape without the differences in coupling patterns that would be expected to emerge from changes in secondary structure. The bleach observed in difference 2D IR spectra appears in HDVE plots as an instantaneous offset of the baseline following the T-jump (Fig. 3). Finally, well-ordered protein aggregates exhibit a characteristic sharp peak near 1615 cm^{-1}, and the peak labeled **c** corresponds to a small amount of protein aggregate present in the sample (see the Supporting Material). The aggregate peak is also observed in high-temperature equilibrium spectra (Fig. 2 *b*, **VI**). $NTL9_{1-39}$ tends to aggregate at temperatures approaching the melting point, however,

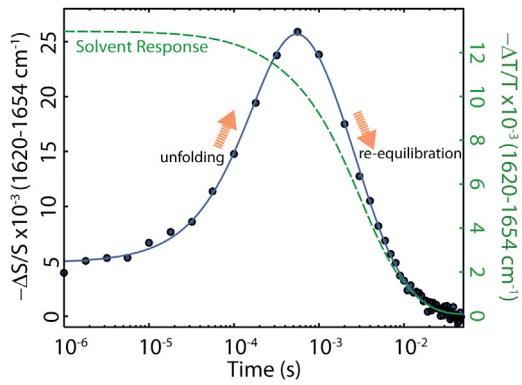

FIGURE 3 Heterodyne-detected vibrational echo response of $NTL9_{1-39}$ following the T-jump response of $NTL9_{1-39}$ at Ti = 60°C. The plot represents the average change in signal between 1620 and 1654 cm^{-1}. The full spectral response is shown in Fig. S5. The solid curve represents a multiexponential fit to the data. The initial rise observed in the 100 ns—1 ms timescale is due to the thermal unfolding of $NTL9_{1-39}$ and the millisecond reequilibration arises from a convolution of the protein response with the solvent reequilibration. The dashed curve shows the exponential relaxation of the solvent. The undistorted protein response is calculated through a numerical deconvolution of these two curves (see the Supporting Material for details). To see this figure in color, go online.

TABLE 1 Equilibrium, folding and unfolding rate constants for $NTL9_{1-39}$ derived from equilibrium FTIR and T-jump HDVE experiments at pH* = 5.6

$T (T_i + \Delta T_{jump})$ [°C]	K_{eq}	$1/k_{obs}$ (μs)	$1/k_{fold}$ (μs)	$1/k_{unfold}$ (μs)
60 + 10	0.939	204	98.8	105.2
70 + 10	0.308	112	26.4	85.6

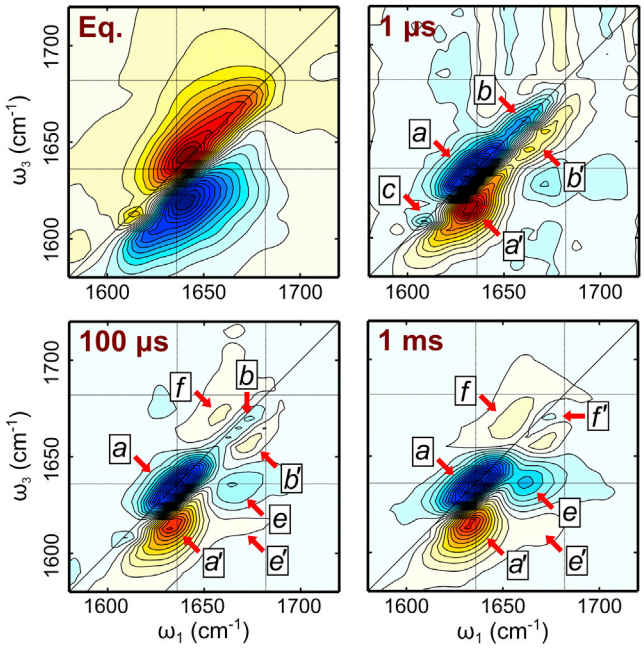

FIGURE 4 Equilibrium (*top right*) absorptive and T-jump difference 2D IR surfaces of $NTL9_{1-39}$ collected in the all parallel polarization at a waiting time of 150 fs. The initial equilibrium temperature is 60°C. T-jump delays are indicated in the individual panels. Positive peaks are shown as red contours and negative peaks in blue. Loss of amplitude appears as a blue feature in the transient spectrum only if the same area appears in red in the equilibrium spectrum, generally above the diagonal, and vice versa. Labels indicate the main spectral features discussed in the text. Because equilibrium peaks appear as a positive/negative doublet due to vibrational anharmonicity, each corresponding peak is labeled with a prime. Spectra are plotted with 30 contours scaled by a hyperbolic sine function to emphasize low-amplitude features. To guide the eye, horizontal and vertical bars are shown at 1636 and 1672 cm^{-1}. To see this figure in color, go online.

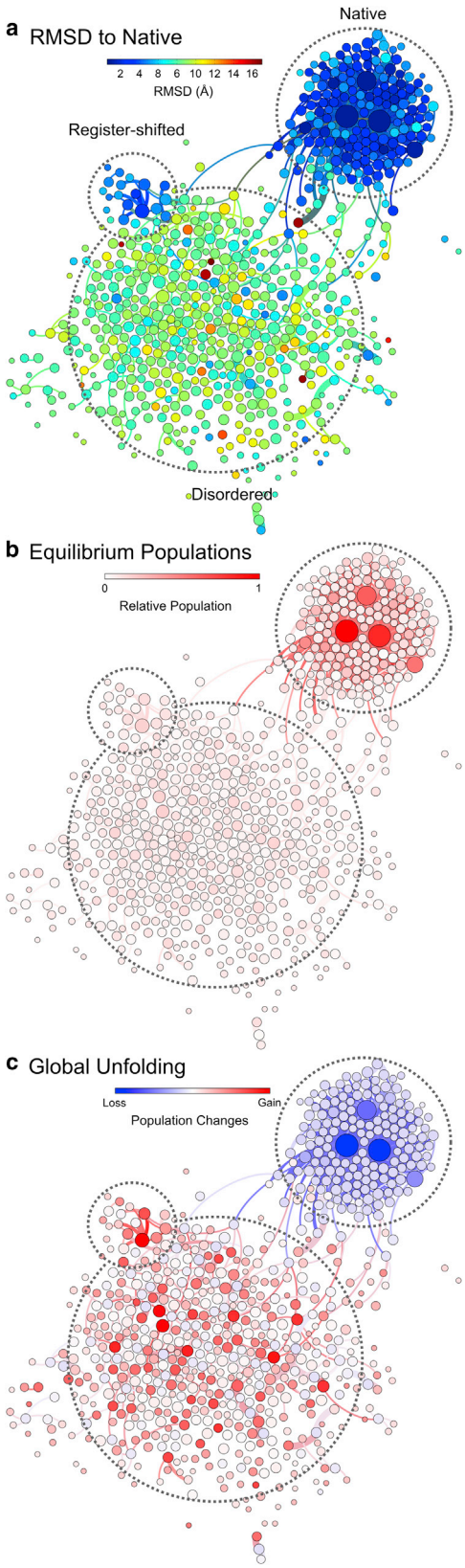

FIGURE 5 Network representations of the MSM for $NTL9_{1-39}$. (*a*) States are shown as nodes color coded by root mean-squared distance of α-carbons to the crystal structure (PDB: 2HBA). The thicknesses of the edges are proportional to the interconversion rates. The network layout is optimized such that states that interconvert rapidly appear at a closer proximity. (*b*) Equilibrium populations as represented by the first eigenvector of the MSM. (*c*) Population changes associated with the slowest transition of the Markov state matrix (18 μs). Negative amplitudes represent loss and positive amplitudes represent gain of population. Color scales for populations are nonlinear to emphasize the low amplitude states. For clarity, only the edges representing fast interconversion rates are plotted. To see this figure in color, go online.

from linear absorption measurements we estimate the aggregate concentration to be <2% of the total protein concentration.

Diagonal bleaches remain present in the 100 μs spectrum with off-diagonal features indicative of changes in secondary structure. The weak positive feature (**f**) grows in around 1660 cm^{-1}, indicating the appearance of a disordered ensemble. Two off-diagonal peaks (**e**, **e**′) indicate the loss of crosspeaks between the low- and high-frequency modes of the sheet. The loss of these peaks is also observed in equilibrium FTIR data (Fig. 1). Changes in diagonal intensities due to loss of β-sheet remain difficult to interpret in 2D IR because the low-frequency mode is overshadowed by the large bleaches (**a**, **a**′) present at short delays, and the high-frequency bleach overlaps with the positive band (**f**) due to the appearance of disordered configurations. Crosspeaks (**e**, **e**′) are fully developed by 1 ms and the disorder peak around 1660 cm^{-1} (**f**) gains intensity, canceling out the diagonal bleaches.

STRUCTURAL AND SPECTRAL MODELING

MSMs

Network layouts of the 727-state MSM for $NTL9_{1-39}$ are shown in Fig. 5. States are represented by nodes whose layout is optimized such that states that interconvert rapidly are placed near each other. The network naturally partitions itself into three main groups: a low-RMSD native state, a high-RMSD disordered group, and a small group of low-RMSD states, which is native-like, but with register-shifted β-strands (67). Native-like states appear as a small cluster connected primarily to the disordered states, whereas the register-shifted state acts as an off-pathway kinetic trap. Inspection of register-shifted structures reveals the hydrogen bonds joining the first and second β-strands are register shifted by one amino acid with respect to the native structure (see the Supporting Material for detailed discussion).

Eigenvector decomposition of the Markov state transition matrix projects out the characteristic transition pathways of the system. The first eigenvector, with a timescale of infinity, gives the stationary equilibrium populations. These are illustrated on the network plot in Fig. 5 *b*, showing that the majority of the population is concentrated into five folded states, whereas the disordered and registry-shifted

states contain small populations. The second eigenvector, illustrated in Fig. 5 *c*, represents the slowest transition: 18 μs at the simulation temperature of 355 K. Higher eigenvectors, with faster transition times show lower barrier transitions in the system.

Physically, the second eigenvector represents the population changes associated with the highest free energy barrier within the system. For a two-state system, such as $NTL9_{1-39}$, the highest barrier corresponds to the global unfolding of the protein, thus the eigenvector populations are negative (loss, blue) for native-like configurations and positive (gain, red) for disordered configurations. The loss of population is dominated by few low-RMSD states, whereas the positive components are distributed over a large number of states, consistent with a two-state picture featuring a rapidly-exchanging heterogeneous unfolded state.

The relaxation rate of the MSM eigenvector associated with the global folding transition is 18 μs, which compares favorably with the 26.4 μs^{-1} rate extracted from the T-jump HDVE measurements (Table 1). Though rates are in agreement, the correspondence should be interpreted with caution. First, the simulations are carried out on the K12M mutant, which destabilizes the denatured state with respect to wild-type, leading to faster folding (44,68). Second, thermodynamic variables do not necessarily correspond to experiments and depend heavily on the choice of force field (69). Nonetheless, the semiquantitative agreement with experiments lends additional validation to the MD simulations and MSM approach.

Spectral simulations

Spectral simulations provide a bridge between protein structures and IR spectra. Fig. 6 shows three simulated spectra for multiple Markov states: Fig. 6 *a* corresponds to states with negative eigenvector amplitudes (population loss, native-like configurations), Fig. 6 *b* shows the positive amplitudes (population gain, disordered configurations), and a difference spectrum of the loss and gain is shown in Fig. 6 *c*. The loss spectrum shows qualitative similarities with the low-temperature experimental equilibrium spectrum shown in Fig. 2. An intense low-frequency peak centered around 1640 cm^{-1} associated with the perpendicular mode of the β-sheet and a high-frequency tail corresponding to the parallel modes are observed. Similar to experiments, crosspeaks appear as off-diagonal ridges. As expected for disordered configurations, the spectrum in Fig. 6 *b* shows a single broad, featureless peak centered around 1660 cm^{-1} spanning much of the amide I′ window. The unfolded spectrum is also in qualitative agreement with the high-temperature experimental spectrum (Fig. 2 *b*).

Finally, and most importantly, the simulated difference spectrum (Fig. 6, *right*) can be compared directly with the measured T-jump difference spectra (Fig. 4). The spectra are in excellent agreement, particularly for the 1 ms T-jump spectrum, where the ensemble is equilibrated at the higher temperature. To highlight the striking similarity between the two spectra, the same labels are used in spectral features in both figures. An intense bleach: (**a**, **a**′) is observed in the 1640 cm^{-1} region, corresponding to the loss of the perpendicular mode of the β-sheet (discussed in the Equilibrium characterization of $NTL9_{1-39}$ section above), the corresponding loss of crosspeaks to the high frequency—but low intensity—parallel mode of the β-sheet (**e**, **e**′) are prominent in both spectra. Finally, the growth of the broad disordered ensemble features, which is observed in the 1670–1700 cm^{-1} region (**f**, **f**′), appears with low intensity in both spectra. The simulated spectrum shows two distinct bands (**f** and $\mathbf{f_2}$), whereas the experiment shows a single broad peak at low frequency, suggesting increased structural heterogeneity in the MSM compared to experiments. It is worth emphasizing that the agreement is achieved with no free model fitting parameters.

Spectroscopic signatures of secondary structure

Interferences between different positive and negative features can lead to unusual, difficult to interpret lineshapes. To further advance the structural interpretation of 2D IR spectra and understand the nature of the lineshapes, in Fig. 7 we show difference spectra, similar to those shown in Fig. 6, but only include residues that are in α-helix (residues 23–33) configurations and β-sheet configurations (residues 1–5, 16–21, 36–38) in the native state. Naturally, the β-sheet difference spectrum bears greater resemblance to the full protein spectrum. This is not surprising because

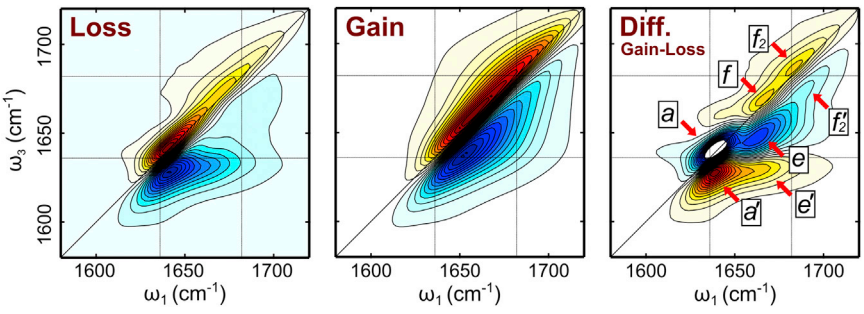

FIGURE 6 Simulated absorptive 2D IR spectra of $NTL9_{1-39}$. Spectra were simulated for each Markov state in the matrix and weighted by the corresponding second-eigenvector. The (*a*) Loss and (b) Gain component spectra are calculated by summing over the states with negative amplitude and positive amplitudes, respectively. The (*c*) difference between these two spectra is shown on the right panel. Horizontal and vertical bars are shown at 1636 and 1672 cm^{-1}. To see this figure in color, go online.

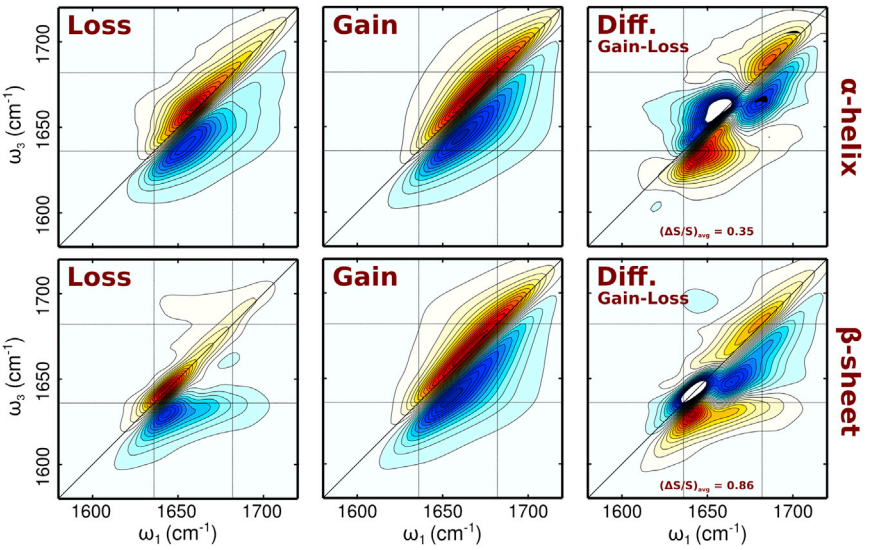

FIGURE 7 Spectral changes associated with different secondary structures (the full protein difference spectrum is shown in Fig. 6). This figure shows difference spectra for residues in (*top*) α-helix and (*bottom*) β-sheet configurations in the folded structure. Contours are normalized and the average amplitudes are indicated in the figure. To see this figure in color, go online.

β-sheets are most distinct compared to disordered configurations, whereas spectral overlap is greater for α-helices (3). The average difference signal is 86% of the total signal in the β-sheet spectrum but only 35% for helices, showing that most of the observed signal arises from the β-sheet response. Nonetheless, the full difference spectrum is in better qualitative agreement with the experimental data, showing that contributions from the helix are also necessary to reproduce the final T-jump spectrum. Finally, because 2D IR spectra represent ensemble measurements, it remains important to understand how structural heterogeneity appears in amide I′ spectra. Toward this end, we simulate spectra using only a subset of states of the MSM. The results shown in Fig. S7, indicate that spectra can be well reproduced with only 11 states (one folded, 10 disordered). This suggests that disordered configurations produce relatively featureless amide I′ spectra, and thus limit the overall structural resolution of the measurement. Because the native state is compact and rigid, the sharp features of the folded spectra are reproduced with only a single folded state (additional discussion is provided in the Supporting Material).

MULTIDIMENSIONAL SPECTROSCOPY AND MARKOV STATE MODELS: IMPLICATIONS FOR TWO-STATE FOLDERS

Two-state folding is primarily defined in thermodynamic terms: two thermodynamic states (configurational ensembles) separated by a large free-energy barrier. This definition is based on a chemical reaction view of protein folding. Although it is an accurate description, it does not easily represent the structural aspects that are crucial to folding, such as residual structure and heterogeneity of the disordered ensemble. To date, most experiments lack the sensitivity required to probe the topology of the folding landscape, thus two-state models remain a popular modeling approach. However, it is important to emphasize that, as demonstrated here, new experimental techniques require models beyond simple two-state descriptions of folding, even for systems that otherwise exhibit typical two-state behavior.

The separation of timescales between the first and subsequent eigenvectors of the transition matrix is consistent with two-state folding behavior. To reduce the complexity of the MSM and more easily interpret the underlying two-state behavior, we examine the contact profile coordinate (Q), which represents the percentage of native-like contacts in the protein. To obtain Q for each structure, the pairwise contact vector (residue pairs within 8 Å in distance between α-carbons) is projected onto the vector for the lowest RMSD to native state for five sample structures of each Markov state. The free energy surface (Fig. 8 *a*) shows two wells separated by a barrier, suggesting that the true folding coordinate maps favorably onto Q. The relative energy of the native to denatured states is ~1.3 kcal/mol, and the free energy barrier to unfold at Q = 0.7 is ~2.35 kcal/mol. It is important to note that because Q is an approximation to the underlying folding coordinate, the barrier along Q represents an upper limit to the true folding barrier.

Because Markov states become naturally partitioned into two groups when mapped along Q, we can estimate the equilibrium constant using the population ratio between the folded (Q >0.8) and denatured (Q <0.5) groups. The extracted equilibrium constant is ~4.2, a value that is not in agreement with experiments. However, it is important to keep in mind that simulations are carried out on the K12M mutant, which destabilizes the solvent-exposed disordered states leading to a higher melting temperature of 81.6°C, compared to 64°C for wild-type (44).

To further describe structural aspects of the two-state behavior, native contact profiles for each secondary structure against the global contact coordinate are shown as

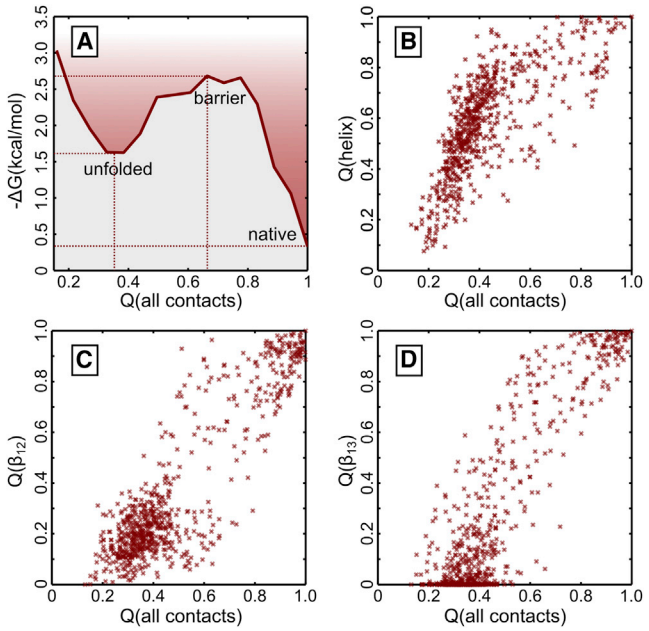

FIGURE 8 (*a*) Single-coordinate folding energy landscape of $NTL9_{1-39}$ extracted from the populations of the Markov states. Native contacts of the (*b*) α-helix, (*c*) first and second β-strands, (*d*) second and third β-strands in relation to the total percentage of native contacts. To see this figure in color, go online.

scatter plots in Fig. 8. The first plot (Fig. 8 *b*) shows a tight correlation between helix formation and global folding for Q <0.6, indicating that a significant number of helical configurations remain present in the disordered state ensemble. The results suggest that native helix formation must occur to reach the transition state. Fig. 8, *b* and *c*, show the native contacts for β_{12} and β_{23} strands against the global native contacts. These coordinates are highly correlated, indicating that sheet formation drives folding in $NTL9_{1-39}$. The results also indicate the transition state contains 60–70% of native β-sheet contacts, results that are in agreement with previous simulations (52). One important caveat of this analysis is that residual helical structure in the disordered ensemble is more likely to contribute to the native contacts because there is a single way to fold a helix—along the chain—whereas native β-sheet contacts require properly aligned strands. So from that perspective, helices exhibit reduced dimensionality of the folding free energy landscape. Overall, the analysis demonstrates that by hiding the complexity of the folding process under a one-dimensional coordinate, simple two-state behavior can be straightforwardly recovered from the MSM, while maintaining the detailed structural insight that is afforded by the MSM.

The folding time ($T_{folding}$) is 98 μs at 70°C (Table 1), from this value we can estimate the free energy barrier to folding ($\Delta G^{\dagger}$) from Kramer's theory. The folding rate can be estimated as (16):

$$T_{folding} = 2\,\pi\tau_{corr}\exp\left(\Delta G^{\dagger}/RT\right).$$

Semiempirical considerations by Eaton and coworkers give a preexponential factor (τ_{corr}) of ~10 us (70). Applying this factor to $NTL9_{1-39}$ we get a $\Delta G^{\dagger}$ of 0.30 kcal/mol, which suggests that the folding process is nearly barrierless, and that the folding barrier may be in good qualitative agreement with the value predicted from MD simulations, particularly since the barrier along Q represents an upper limit to the molecular barrier (~1 kcal/mol, Fig. 8).

It is worth briefly discussing the implications of two-state folding models and the projection of full free energy space (as represented by the MSM) onto a one-dimensional coordinate. As a two-state folder, $NTL9_{1-39}$ is an excellent test case for exploring the relationship between full-dimensional and two-state models. We note that in our simulations the RMSD and Q-values are well correlated with a coefficient of 0.87, so qualitatively, for the purposes of discussion, the RMSD-color-coded MSM plot in Fig. 5 *a* can be interpreted as a Q-value. The figure shows multiple pathways for rapid interconversion between native-like states, within the two-state this can be basically thought of as intrawell relaxation with minimal changes in global structure. Showing that the lifetimes of individual folded microstates are relatively low, this is mainly related to heterogeneity and flexibility of the native structure. In addition, the high degree of connectivity and fast rates of the native Markov states show that these conformations behave as kinetic hubs, consistent with previous observations (71). The one-dimensional representation implies that coordinates orthogonal to Q exhibit fast dynamics, and are thus averaged out. However, the kinetic hub topology of the MSM implies that the picture involving a single barrier or transition state ensemble may be too simplistic (72–74). The results indicate that the global folding coordinate must be chosen carefully to encapsulate the maximum amount of microscopic detail present in the MSM.

Since the first simulations involving polymer chains, much work has been put into developing a conceptual understanding of folding from numerical trajectories (13,17,75,76). Ideas such as the funnel-like topology of the folding landscape or the glassy dynamics near the minimum emerged from these early efforts (18,74,77). On the other hand, the wealth of data provided by millisecond-long MD trajectories, which have only become attainable in recent years, calls for a redevelopment of coordinate mapping schemes that can provide an intuitive view of the folding process while retaining the structural details present in the simulations (19,21,78). More generally, the recent abundance of simulations and experimental data provides an excellent opportunity to revisit the established models of protein folding.

CONCLUDING REMARKS

To date, MD simulations have been largely validated at the level of equilibrium experimental structures and folding

rates. This work showcases the use of T-jump amide I′ 2D IR spectroscopy as a structurally sensitive technique to study protein folding with nanosecond resolution over several decades in time. Unfolding timescales extracted from simulations are in excellent agreement with the measured response times of $NTL9_{1-39}$. Spectral simulations based on the full 727-state MSM are in semiquantitative agreement with T-jump transient 2D IR. Of importance, residual β-sheet contents in the denatured state ensemble show agreement between the experiment and simulation suggesting minimal bias in force field used.

The analysis in this work emphasizes the overall agreement between the experiment and MSM for describing global conformational changes of $NTL9_{1-39}$ on folding or unfolding. To work toward a more comprehensive description, future studies will aim to spectroscopically distinguish conformational states within the folded and unfolded while addressing similarities between folded and unfolded ensembles and the exchange time between configurations. For these purposes, a strategic selection of site-specific isotope labels in the $NTL9_{1-39}$ backbone should be able to reveal conformers that vary by local hydrogen bonding contacts, as illustrated recently for β-hairpin folding (2). Thus, we believe that the combination of transient 2D IR spectroscopy with state-of-the-art MD simulations can be used to provide increasingly rich structural insight into the free energy topography of proteins and serve to validate new modeling methods.

SUPPORTING MATERIAL

Seven figures, supporting data, and references (79-80) are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00184-2.

The authors thank Prof. Daniel P. Raleigh for advice on synthesis and characterization of protein samples. The authors thank D. E. Shaw Research for providing the raw molecular dynamics trajectories.

The project was supported by the National Science Foundation (CHE-0911107), MIT Laser Biomedical Research Center (P41-EB015871), Agilent Technologies, and a grant from the National Eye Institute (PN2EY016525). C.R.B. was supported by a Ruth L. Kirschstein National Research Service Award (F32GM105104). Y.S.L. was supported by the Bio-X postdoctoral fellowship at Stanford University. K.A.B. was funded by a Stanford graduate fellowship. The authors acknowledge the following award for providing computing resources that have contributed to the research results reported within this paper: MRI-R2: Acquisition of a Hybrid CPU/GPU and Visualization Cluster for Multidisciplinary Studies in Transport Physics with Uncertainty Quantification. This award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5).

REFERENCES

1. Chung, H. S., K. McHale, …, W. A. Eaton. 2012. Single-molecule fluorescence experiments determine protein folding transition path times. *Science.* 335:981–984.
2. 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.
3. Ganim, Z., H. S. Chung, …, A. Tokmakoff. 2008. Amide I two-dimensional infrared spectroscopy of proteins. *Acc. Chem. Res.* 41:432–441.
4. Serrano, A., M. Waegele, and F. Gai. 2012. Spectroscopic studies of protein folding: linear and nonlinear methods. *Protein Sci.* 21:157–170.
5. Greenfield, N. J. 2006. Analysis of the kinetics of folding of proteins and peptides using circular dichroism. *Nat. Protoc.* 1:2891–2899.
6. Royer, C. A. 2006. Probing protein folding and conformational transitions with fluorescence. *Chem. Rev.* 106:1769–1784.
7. Lange, R., and C. Balny. 2002. UV-visible derivative spectroscopy under high pressure. *Biochim. Biophys. Acta.* 1595:80–93.
8. Fabian, H., and D. Naumann. 2004. Methods to study protein folding by stopped-flow FT-IR. *Methods.* 34:28–40.
9. Schuler, B., and W. A. Eaton. 2008. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* 18:16–26.
10. Reference deleted in proof.
11. Schuler, B. 2005. Single-molecule fluorescence spectroscopy of protein folding. *Chemphyschem.* 6:1206–1220.
12. Dudko, O. K., G. Hummer, and A. Szabo. 2008. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proc. Natl. Acad. Sci. USA.* 105:15755–15760.
13. 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.
14. Dill, K. A. 1999. Polymer principles and protein folding. *Protein Sci.* 8:1166–1180.
15. Naganathan, A. N., U. Doshi, …, V. Muñoz. 2006. Dynamics, energetics, and structure in protein folding. *Biochemistry.* 45:8466–8475.
16. Kubelka, J., J. Hofrichter, and W. A. Eaton. 2004. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* 14:76–88.
17. Socci, N. D., J. N. Onuchic, and P. G. Wolynes. 1998. Protein folding mechanisms and the multidimensional folding funnel. *Proteins.* 32:136–158.
18. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.
19. Dill, K., S. Ozkan, M. Shell, and T. Weikl. 2008. The protein folding problem. *Ann. Rev. Biophys.* 37:289–316.
20. Gruebele, M. 2002. Protein folding: the free energy surface. *Curr. Opin. Struct. Biol.* 12:161–168.
21. Lindorff-Larsen, K., S. Piana, …, D. E. Shaw. 2011. How fast-folding proteins fold. *Science.* 334:517–520.
22. Christian, R. S., and S. P. Vijay. 2013. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* 9:2000–2009.
23. Beauchamp, K. A., R. McGibbon, …, V. S. Pande. 2012. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. USA.* 109:17807–17813.
24. Dickson, A., and C. L. Brooks, 3rd. 2013. Native states of fast-folding proteins are kinetic traps. *J. Am. Chem. Soc.* 135:4729–4734.
25. Bowman, G. R., V. A. Voelz, and V. S. Pande. 2011. Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* 21:4–11.
26. Barth, A. 2007. Infrared spectroscopy of proteins. *Biochim. Biophys. Acta.* 1767:1073–1101.
27. Barth, A., and C. Zscherp. 2002. What vibrations tell us about proteins. *Q. Rev. Biophys.* 35:369–430.
28. Baiz, C., M. Reppert, and A. Tokmakoff. 2012. Amide I two-dimensional infrared spectroscopy: methods for visualizing the vibrational structure of large proteins. *J. Phys. Chem. A.* 117:5955–5961.
29. Ganim, Z., and A. Tokmakoff. 2006. Spectral signatures of heterogeneous protein ensembles revealed by MD Simulations of 2DIR spectra. *Biophys. J.* 91:2636–2646.
30. Hamm, P., and M. T. Zanni. 2011. Concepts and Methods of 2D Infrared Spectroscopy. Cambridge University Press, Cambridge; New York.

31. 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.
32. Hochstrasser, R. M. 2007. Two-dimensional spectroscopy at infrared and optical frequencies. *Proc. Natl. Acad. Sci. USA.* 104:14190–14196.
33. Middleton, C. T., A. M. Woys, …, M. T. Zanni. 2010. Residue-specific structural kinetics of proteins through the union of isotope labeling, mid-IR pulse shaping, and coherent 2D IR spectroscopy. *Methods.* 52:12–22.
34. Cho, M. 2008. Coherent two-dimensional optical spectroscopy. *Chem. Rev.* 108:1331–1418.
35. Reference deleted in proof.
36. Wang, L., C. T. Middleton, …, J. L. Skinner. 2011. Development and validation of transferable amide I vibrational frequency maps for peptides. *J. Phys. Chem. B.* 115:3713–3724.
37. Smith, A. W., J. Lessing, …, J. Knoester. 2010. Melting of a beta-hairpin peptide using isotope-edited 2D IR spectroscopy and simulations. *J. Phys. Chem. B.* 114:10913–10924.
38. Shaw, D. E., M. M. Deneroff, …, J. C. Chao. 2007. Anton, a special-purpose machine for molecular dynamics simulation. *ACM.* 35:1–12.
39. Shaw, D. E., R. O. Dror, …, K. J. Bowers. 2009. Millisecond-Scale Molecular Dynamics Simulations on Anton. *IEEE.* 1–11.
40. 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.
41. Prinz, J. H., B. Keller, and F. Noé. 2011. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.* 13:16912–16927.
42. Prigozhin, M. B., and M. Gruebele. 2013. Microsecond folding experiments and simulations: a match is made. *Phys. Chem. Chem. Phys.* 15:3372–3388.
43. Chung, H. S., Z. Ganim, …, A. Tokmakoff. 2007. Transient 2D IR spectroscopy of ubiquitin unfolding dynamics. *Proc. Natl. Acad. Sci. USA.* 104:14237–14242.
44. 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.
45. Piana, S., K. Lindorff-Larsen, and D. E. Shaw. 2011. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100:L47–L49.
46. Shaw, D. E., P. Maragakis, …, W. Wriggers. 2010. Atomic-level characterization of the structural dynamics of proteins. *Science.* 330: 341–346.
47. Chung, H. S., M. Khalil, …, A. Tokmakoff. 2007. Transient two-dimensional IR spectrometer for probing nanosecond temperature-jump kinetics. *Rev. Sci. Instrum.* 78:063101-1–063101-10.
48. Jones, K. C., Z. Ganim, and A. Tokmakoff. 2009. Heterodyne-detected dispersed vibrational echo spectroscopy. *J. Phys. Chem. A.* 113:14060–14066.
49. Jones, K., Z. Ganim, …, A. Tokmakoff. 2012. Transient two-dimensional spectroscopy with linear absorption corrections applied to temperature-jump two-dimensional infrared. *JOSA B.* 29:118–129.
50. 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.
51. Gonzalez, T. F. 1985. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.* 38:293–306.
52. 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.
53. Masucci, A. P., A. Kalampokis, …, E. Hernández-García. 2011. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PLoS ONE.* 6:e17333.
54. Jansen, T. L., A. G. Dijkstra, …, J. Knoester. 2006. Modeling the amide I bands of small peptides. *J. Chem. Phys.* 125:44312-1–44312-9.
55. Jansen, T. L., 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.* 124:044502-1–044502-11.
56. Van Der Spoel, D., E. Lindahl, …, H. J. C. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.
57. 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.
58. Becktel, W. J., and J. A. Schellman. 1987. Protein stability curves. *Biopolymers.* 26:1859–1877.
59. Cheatum, C. M., A. Tokmakoff, and J. Knoester. 2004. Signatures of beta-sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J. Chem. Phys.* 120:8201–8215.
60. Smith, A. W., C. M. Cheatum, …, A. Tokmakoff. 2004. Two-dimensional infrared spectroscopy of beta-sheets and hairpins. *Biophys. J.* 86:619a–619a.
61. Smith, A. W., H. S. Chung, …, A. Tokmakoff. 2005. Residual native structure in a thermally denatured beta-hairpin. *J. Phys. Chem. B.* 109:17025–17027.
62. Smith, A. W., and A. Tokmakoff. 2007. Probing local structural events in beta-hairpin unfolding with transient nonlinear infrared spectroscopy. *Angew. Chem. Int. Ed. Engl.* 46:7984–7987.
63. Anil, B., Y. Li, …, D. P. Raleigh. 2006. The unfolded state of NTL9 is compact in the absence of denaturant. *Biochemistry.* 45:10110–10116.
64. Meng, W., B. Luan, …, D. Raleigh. 2013. The denatured state ensemble contains significant local and long-range structure under native condition: analysis of the N-terminal domain of the ribosomal protein L9. *Biochemistry.* 52:2662–2671.
65. Joosten, R. P., T. A. te Beek, …, G. Vriend. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* 39 (Database issue):D411–D419.
66. Chung, H. S., and A. Tokmakoff. 2006. Visualization and characterization of the infrared active amide I vibrations of proteins. *J. Phys. Chem. B.* 110:2888–2898.
67. Evans, M. R., and K. H. Gardner. 2009. Slow transition between two β-strand registers is dictated by protein unfolding. *J. Am. Chem. Soc.* 131:11306–11307.
68. Cho, J.-H., S. Sato, and D. P. Raleigh. 2004. Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state. *J. Mol. Biol.* 338:827–837.
69. Ponder, J. W., and D. A. Case. 2003. Force fields for protein simulations. *Adv. Protein Chem.* 66:27–85.
70. Muñoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA.* 96:11311–11316.
71. Bowman, G. R., and V. S. Pande. 2010. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. USA.* 107:10890–10895.
72. Sosnick, T. R., and D. Barrick. 2011. The folding of single domain proteins—have we reached a consensus? *Curr. Opin. Struct. Biol.* 21:12–24.
73. Lindberg, M. O., and M. Oliveberg. 2007. Malleability of protein folding pathways: a simple reason for complex behaviour. *Curr. Opin. Struct. Biol.* 17:21–29.
74. Dill, K. A., and H. S. Chan. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10–19.
75. Dill, K., S. Bromberg, …, H. Chan. 1995. Principles of protein folding–a perspective from simple exact models. *Protein Sci.* 4:561–602.
76. Gruebele, M. 1999. The fast protein folding problem. *Annu. Rev. Phys. Chem.* 50:485–516.

77. Dill, K. 1999. Polymer principles and protein folding. *Protein Sci.* 8:1166–1180.

78. Ma, H., and M. Gruebele. 2006. Low barrier kinetics: dependence on observables and free energy surface. *J. Comput. Chem.* 27:125–134.

79. Baiz, C. R., R. McCanne, and K. J. Kubarych. 2010. Transient vibrational echo versus transient absorption spectroscopy: a direct experimental and theoretical comparison. *Appl. Spectorsc.* 64: 1037–1044.

80. Nagarajan, K. S., H. Taskent-Sezgin, …, R. Dyer. 2011. Differential ordering of the protein backbone and side chains during protein folding revealed by site-specific recombinant infrared probes. *J. Am. Chem. Soc.* 133:20335–20340.

SUPPORTING MATERIAL

A molecular interpretation of 2D IR protein folding experiments with Markov state models

Carlos R. Baiz[†,**], Yu-Shan Lin[*,♯], Chunte Sam Peng[†], Kyle A. Beauchamp[‡], Vincent A. Voelz[¶], Vijay S. Pande[*,‡,§], and Andrei Tokmakoff[†,**]

[*]Department of Chemistry, [‡]Biophysics Program, [§]Department of Structural Biology, Stanford University, Stanford, CA, USA; [†]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA; [¶]Department of Chemistry, Temple University, Philadelphia, PA, USA; [♯]Current address: Department of Chemistry, Tufts University, Medford, MA, USA, **Current address: Department of Chemistry and James Franck Institute, University of Chicago, Chicago, IL, USA

S1. Markov State Model

1.1 Characterization of MSM eigenvectors

Similar to a principal component analysis, the magnitude of each eigenvector component of the Markov transition matrix represents changes in populations associated with the principal transitions of the system.(1) Eigenvalues give the rate constant of the corresponding transition: the first eigenmode (Figure S1, top left) describes the equilibrium population with a timescale of infinity, and the second eigenvector represents the changes in populations associated with the slowest structural interconversion in the system, namely the global unfolding. The most populated state has a C_α-RMSD of ~0.7 Å to the crystal structure (PDB 2HBA). By sorting the states based on their RMSD to experiment (Figure S1, top center), we find that the states with large equilibrium population tend to have small RMSD to the experimental native structure.

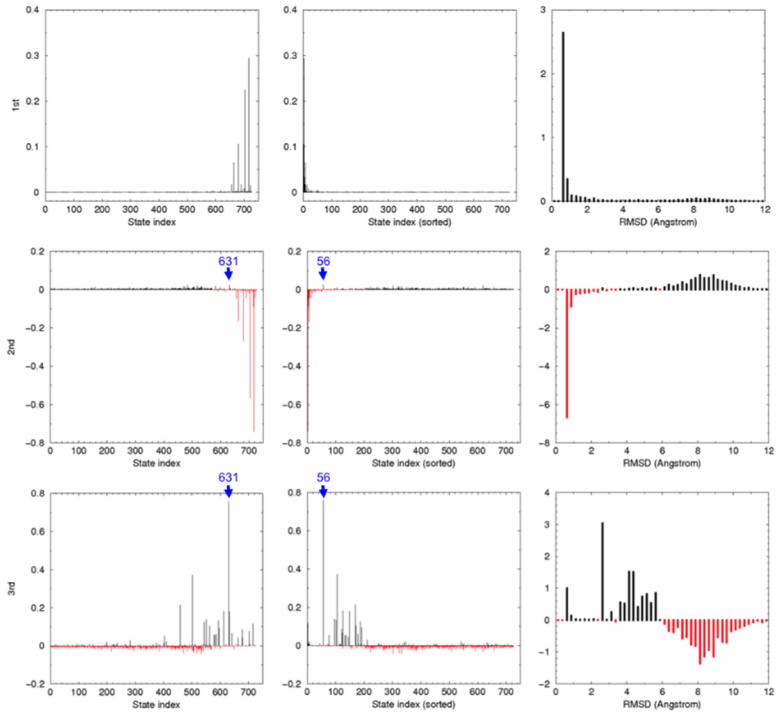

Figure S1: The first (top row), second (middle row), and third (bottom row) eigenmodes of the MSM. The state index in the left column is from the original MSM construction; the states are sorted by Cα-RMSD to the experimental native structure (PDB 2HBA) from low to high in the center column. The blue arrows indicate the state index for the near-native, register-shifted state (631 before sorting and 56 after sorting). The right column is the histogram of RMSD to experiment for the eigenmodes.

1.2 Register-shifted kinetic trap

While the global folding remains the slowest process (~18 μs at the simulation temperature of 355 K), the second slowest process (~2.3 μs, Figure S1 bottom row) describes the forming and breaking of a register-shifted conformation (Figure S2, right). We find that although this state appears native-like, its β strands display a one-residue register-shift. For instance, in the most populated state (Figure S2, left), the carbonyl group of K19 hydrogen bonds with the amide hydrogen of V3; in the register-shifted state (Figure S2, right) the carbonyl group of K19 hydrogen bonds with the amide hydrogen of I4. We note that this 1-residue shift in register is not a mere up and down "shift" between β strands, due to the staggered orientations of the amide groups within a β strand, an additional flip by 180° is required to recover the original register. Therefore, to refold this seemingly near-native register-shifted structure into the native register, rather than merely "shift" the strands by one residue the peptide must break up the shifted β strands and then form the correct register. Such a slow transition between two β-strand registers was recently reported in a mutated PAS-B domain of ARNT protein, where a conformation with a β-strand register shift from the native register is present.(2) In the case of the PAS-B domain, the interconversion between the native and the shifted registers is slow enough that the two registers can be separated by ion exchange chromatography and the conversion kinetics studied by NMR spectroscopy. The shifted and native registers were suggested to interconvert through unfolding and the presence and the control of such register interconversion may be physiologically relevant. Our MSM analysis identifies, *a priori*, a similarly register-shifted kinetic trap in $NTL9_{1-39}$ that is difficult to discover by more traditional projection analyses of the MD trajectory data.

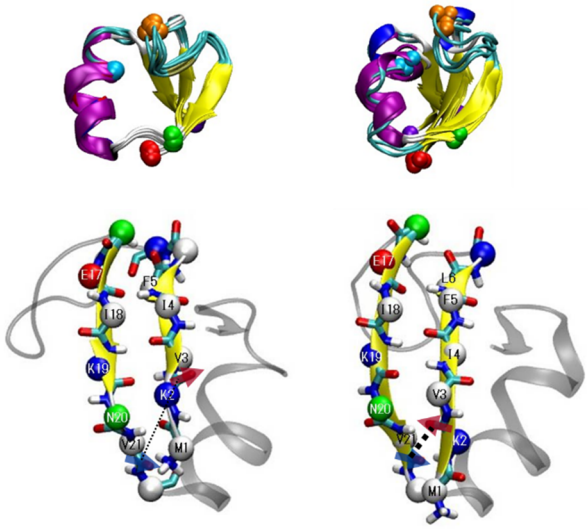

Figure S2: The structure of the most populated state, left, Cα-RMSD of ~0.7 Å to the experimental native structure [PDB: 2HBA], and the near-native register-shifted state, right, Cα-RMSD of ~2.7 Å to the experimental native structure. Top: Five random configurations selected from each state; bottom: view from the back of the β strands to demonstrate the register.

S2. $NTL9_{1-39}$ Equilibrium Spectra – Thermal Denaturation Experiments

The thermal denaturation curve shown in Figure 1c was fit to a sigmoidal function of the type

$$P(T) = A/(1+exp(-s(T-T_0))) + \mathrm{B}$$

Where A and B are coefficients that account for the amplitude and offset of the signal respectively, s is a stretching factor and T_0 is approximately the melting temperature. The curve provides an excellent representation of the curve with RMS residuals of less than 0.3%.

The folded population curve was then constructed by setting A and B to 1 and zero respectively. These values are then turned into a $\Delta G_{F\to U}$ and fitted to the thermodynamic relation described in section 3.1 (main text). The thermodynamic errors were estimated by repeating the $\Delta G_{F\to U}$ calculation and curve fitting while randomly sampling s and T_0 from a normal distribution with the standard deviations given by the confidence intervals above. This procedure was repeated 100 times and the RMS errors are reported as the fitting errors in Section 3.1 of the main text.

It is worth pointing out the fact that $NTL9_{1\text{-}39}$ is prone to aggregation, particularly at high temperatures, and because of the relatively high concentrations required for IR spectroscopy (~2.5 mM), care must be taken to ensure that aggregation is minimized during the measurements. Fortunately aggregates exhibit a sharp peak at 1612-1615 cm^{-1}, which can be used to estimate the amount of aggregate present in the sample. Figure S3 shows an FTIR spectrum collected at 92 ºC following a temperature ramp. The spectrum shows a shoulder appearing in the 1615 cm^{-1} region. To estimate the amount of aggregate we fit the spectrum to a sum of multiple Gaussians (blue curve), and calculate the fraction of total intensity corresponding to the aggregate peak. This method yields a total aggregate concentration of 0.76%. In 2D IR spectra the non-linear character of the signal significantly "amplifies" the aggregate peak in comparison to the main band, making the method even more sensitive to aggregates and enabling us to monitor aggregation in real-time during the measurements. We observed significant aggregation in several samples and the data for these measurements was discarded; only data with an aggregate concentration below ~2%, was analyzed. To slow the aggregation process, protein concentrations were kept below 10 mg/mL and CaF2 windows were replaced after each data acquisition run, however, the spectrally distinct sharp features of the protein aggregates make them easy to distinguish from the main protein response. Finally, it is worth pointing out that while IR spectroscopy is very sensitive to well-structured aggregates with long-range order, the technique cannot reliably distinguish between protein monomers and oligomers that may be present in solution.

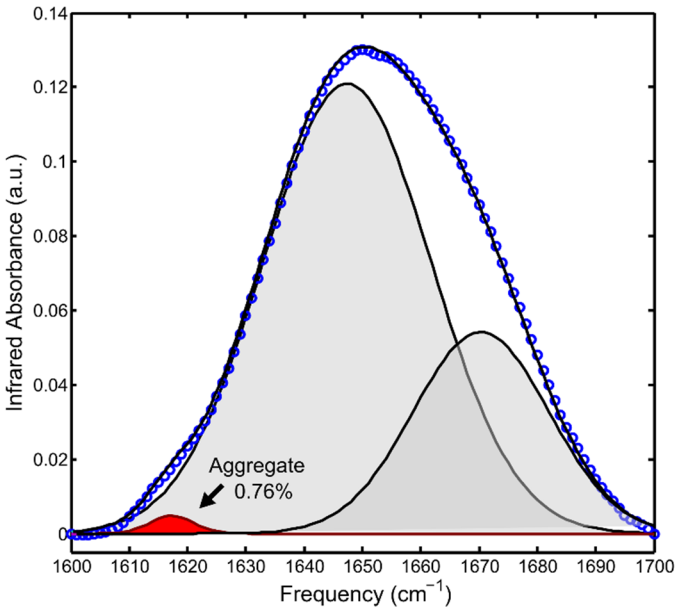

Figure S3: FTIR spectrum of $NTL9_{1\text{-}39}$ collected at 90 ºC. The black curves represent the individual Gaussians along with the sum and the red peak indicates the absorption peak corresponding to the aggregate.

Temperature-dependent 2D IR spectra of $NTL9_{1-39}$ (not shown) are analyzed for secondary structure content using a singular value decomposition based on a procedure described elsewhere.(3) In brief, the spectra projected along a 'pure' β-sheet spectrum derived from a set of sixteen well-characterized proteins. The amplitude of this projection is directly proportional to the number of residues in β -sheet conformation. At low-temperatures, approximately 25% of the residues are in β -sheet configurations, in good agreement with a structural analysis of the crystal structure which shows that 28% of the 39 residues compose the three-strand β -sheet. Two-dimensional spectra were collected at 10 ºC intervals from 0 ºC to 90 ºC. At high temperatures, β-sheet contents is observed to decrease to a ~10% (Figure S4). The temperature-dependence of the β -sheet curve mirrors the thermal denaturation curve of $NTL9_{1-39}$ (see main text).

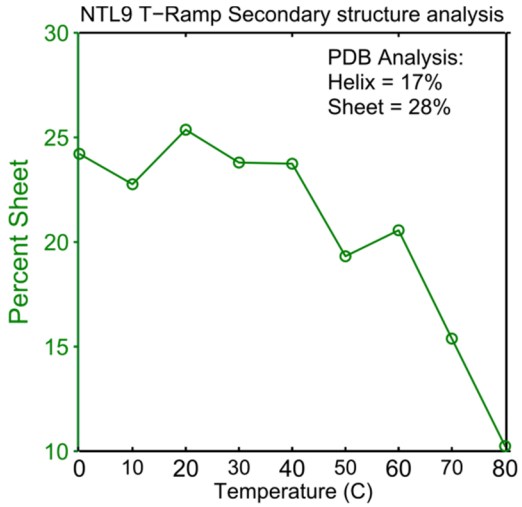

Figure S4: Structural decomposition of 2D IR spectra of $NTL9_{1-39}$ as a function of temperature.

S3. Temperature-jump Heterodyned-detected Vibrational Echo (HDVE)

Temperature jump HDVE spectra were collected in order to measure the nonequilibrium response of $NTL9_{1-39}$. Detailed descriptions of the data processing steps are provided in previous publications.(4) In order to remove possible artifacts due to phasing errors, the absolute value of the complex HDVE signal is used to extract the temporal response. Unlike homodyne-detected signals, heterodyned signals are linear with respect to the sample concentration.(5, 6) Figure S4 shows a set of transient HDVE spectra for $NTL9_{1-39}$ at Ti=60 ºC along with the temporal response of the negative part of the signal. The solvent re-equilibration is monitored by recording the changes in transmission of the reference. A stretched exponential fit to the transmission, $\Delta T(t) = exp\left[-(t/\tau)^{\beta}\right]$, gives a time constant of 3.52 ± 0.1 ms, and a β factor of 0.77±0.03 for an initial temperature of 60 ˚C (the values are nearly identical for T_i=70 ˚C). The measured sample response is a convolution of the response of the protein and the solvent relaxation. In order to extract the undistorted sample response, a deconvolution is carried out as follows: *1*. The raw T-jump data is interpolated to logarithmically spaced points between 1 μs and 50 ms. *2*. The data is fit to a sum of three stretched exponential functions to obtain a smooth fit. An example of the data and best fit line are shown in Figure 3 of the main text. *3*. The solvent transmission curves are interpolated to the same time points and fit to a stretched exponential. *4*. The fit functions are numerically deconvolved using linearly spaced points. *5*. The deconvolved function is re-interpolated back to the original logarithmically spaced time points. *6*. Finally, the deconvolved response is fit to a single non-stretched exponential. All data analysis was carried out using the MATLAB package of programs (R2011a, The Mathworks, Natick, MA).

Figure S5 shows the response of the protein at two different initial temperatures (Ti=60 ˚C and 70 ˚C). The data is discussed in the main text. Exponential fits of deconvolved response yield time constants of 204 µs (60 ˚C) and 112 µs (70 ˚C) with uncertainties of approximately 9 µs. Using the equilibrium thermodynamic parameters measured by FTIR (see main text) we can extract the folding and unfolding rate constants as shown in Table 1 in the main text.

It is important to point out that the recently reported(7) value of 152 µs at T_i=60 ºC is somewhat shorter than our measured response of 204 µs, but the reported value does not properly deconvolve the response from the solvent re-equilibration as described here.

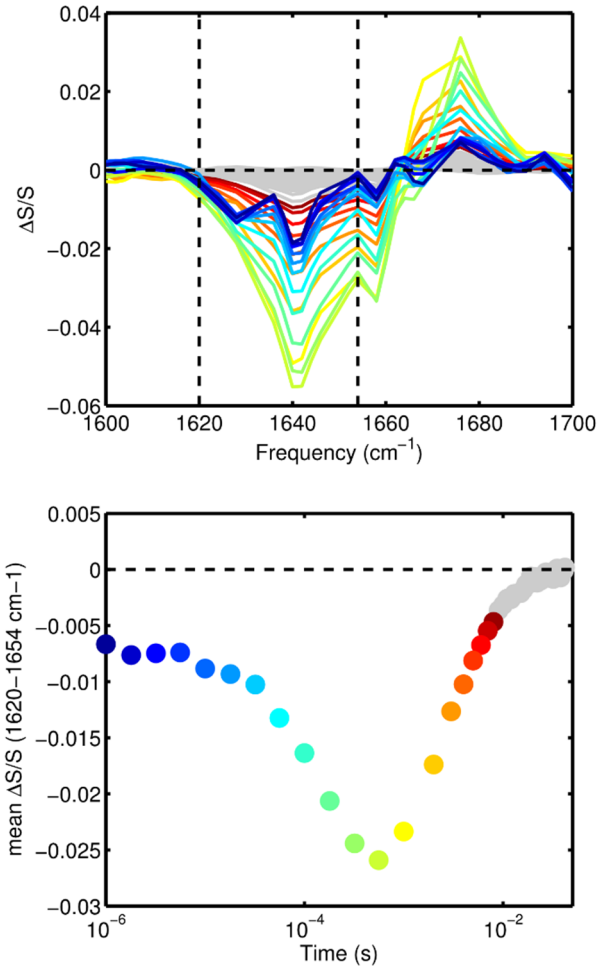

Figure S5: Temperature-jump HDVE spectral response of $NTL9_{1-39}$ at Ti=60 ºC along with the mean temporal response in the low-frequency region (1620-1654 cm^{-1}). The temporal response along with the exponential fit is also plotted in Figure 3 (main text).

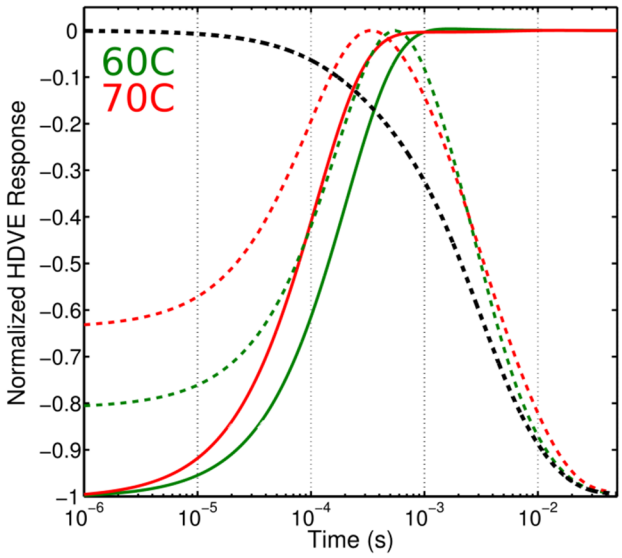

Figure S6: Temperature-jump HDVE temporal response of $NTL9_{1-39}$ at Ti=60 °C (green curves) and Ti=70 °C (red curves). The curves correspond to the average response between 1620 and 1654 cm^{-1}. The fitted water re-equilibration curve at 60 °C is shown by the black curve. Dashed curves represent fits to the measured response and solid curves represent the deconvolved protein response.

S4. Spectral Signatures of Structural Heterogeneity

The simulations presented here represent ensemble averages over 727 Markov states. Here we explore the sensitivity of 2D IR to heterogeneity by reducing the number of states used for the simulation. Figure S7 shows a set of spectra simulated using only 11 states: 1 folded (state 717), and ten disordered states showing maximum positive second-eigenvector amplitude. Spectra simulated with fewer than ten unfolded states are in poorer agreement with experiments. The purpose of the analysis is to test the structural resolution limitations of 2D IR spectroscopy. Not surprisingly, the spectra are well-reproduced with only 11 dominant states. The interpretation is two-fold: Firstly, though the spectra of the folded state are relatively sharp, the structural heterogeneity of the folded configuration is small, therefore well represented by a single state. Secondly, the disordered states are heterogeneous but the amide I' bands are broad and featureless, thus only a relatively small set of structures is needed in order to reproduce the spectra. Coarse graining of the Markov states in combination with maximum entropy methods can be used to further refine the structures and reduce the number of states needed to reproduce experiments. Finally, it is important to point out that the experiments presented here only rely on the main amide I' band. Strategic isotope labeling of residues can be used to probe local contacts to further explore the equilibrium structure and folding mechanism of $NTL9_{1-39}$.

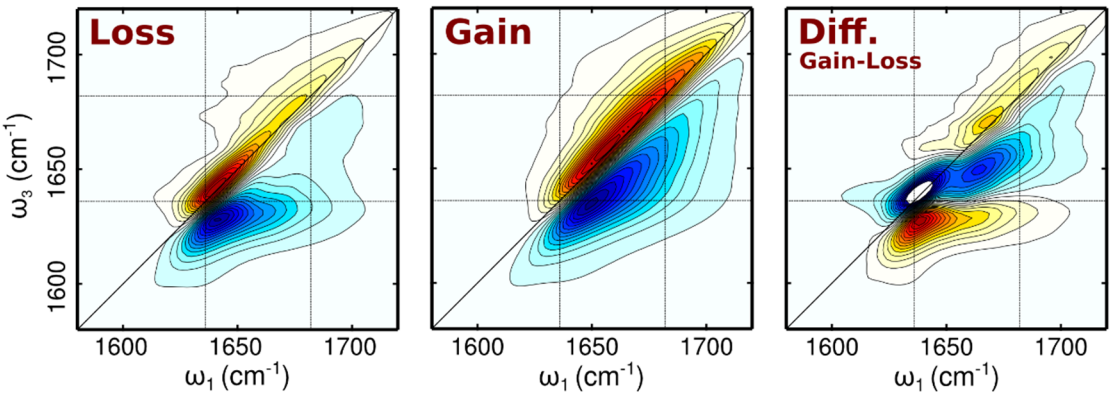

Figure S7: Spectra simulated using a reduced number of Markov states (1 folded, 10 disordered). See Figure 6 in the main text.

Supporting References

1. Prinz, J. H., B. Keller, and F. Noe. 2011. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. Phys. Chem. Chem. Phys. 13:16912-16927.
2. Evans, M. R., and K. H. Gardner. 2009. Slow transition between two β-strand registers is dictated by protein unfolding. J. Am. Chem. Soc. 131:11306-11307.
3. Baiz, C., C. Peng, M. Reppert, K. Jones, and A. Tokmakoff. 2012. Coherent two-dimensional infrared spectroscopy: quantitative analysis of protein secondary structure in solution. The Analyst 137:1793-1799.
4. Jones, K., Z. Ganim, C. Peng, and A. Tokmakoff. 2012. Transient two-dimensional spectroscopy with linear absorption corrections applied to temperature-jump two-dimensional infrared. JOSA B.
5. Jones, K. C., Z. Ganim, and A. Tokmakoff. 2009. Heterodyne-Detected Dispersed Vibrational Echo Spectroscopy. J Phys Chem A 113:14060-14066.
6. Baiz, C. R., R. McCanne, and K. J. Kubarych. 2010. Transient vibrational echo versus transient absorption spectroscopy: a direct experimental and theoretical comparison. Appl Spectrosc 64:1037-1044.
7. Nagarajan, S., H. Taskent-Sezgin, D. Parul, I. Carrico, D. Raleigh, and R. Dyer. 2011. Differential ordering of the protein backbone and side chains during protein folding revealed by site-specific recombinant infrared probes. J. Am. Chem. Soc. 133:20335-20340.