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.

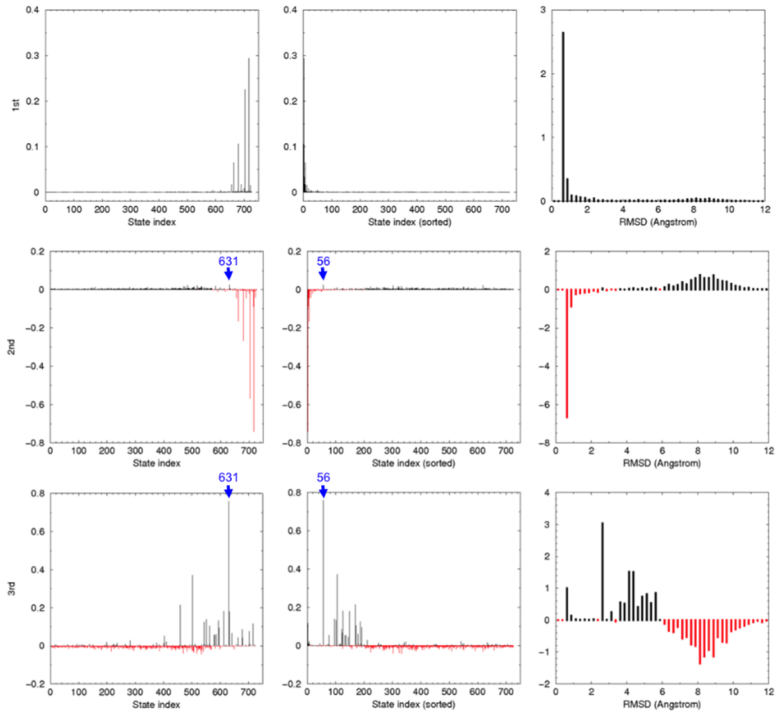

Figure S1: The first (top row), second (middle row), and third (bottom row) eigenmodes of the MSM. The state index in the left column is from the original MSM construction; the states are sorted by Cα-RMSD to the experimental native structure (PDB 2HBA) from low to high in the center column. The blue 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.

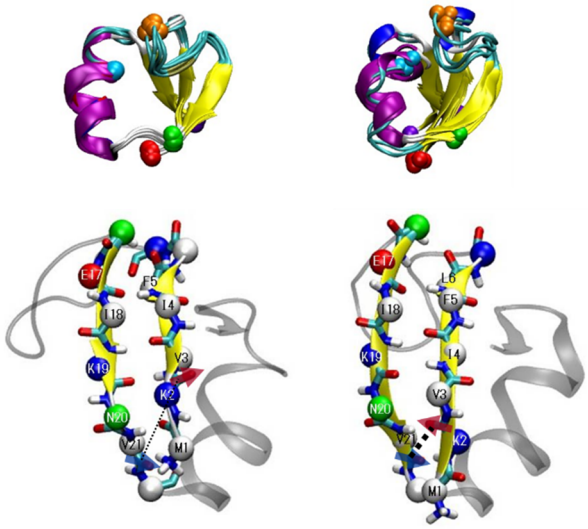

Figure S2: The structure of the most populated state, left, Cα-RMSD of ~0.7 Å to the experimental native structure [PDB: 2HBA], and the near-native register-shifted state, right, Cα-RMSD of ~2.7 Å to the experimental native structure. Top: Five random configurations selected from each state; bottom: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

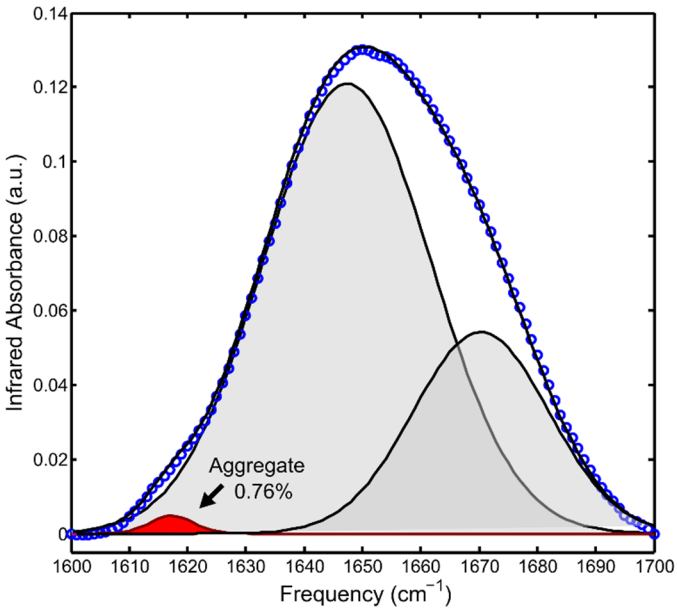

Figure S3: FTIR spectrum of $NTL9_{1\text{-}39}$ collected at 90 ºC. The black curves represent the individual Gaussians along with the sum and the red peak indicates the absorption peak corresponding to the aggregate.

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).

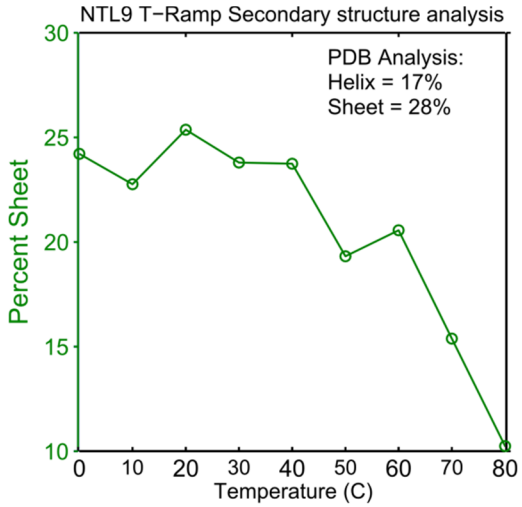

Figure S4: Structural decomposition of 2D IR spectra of $NTL9_{1-39}$ as a function of temperature.

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.

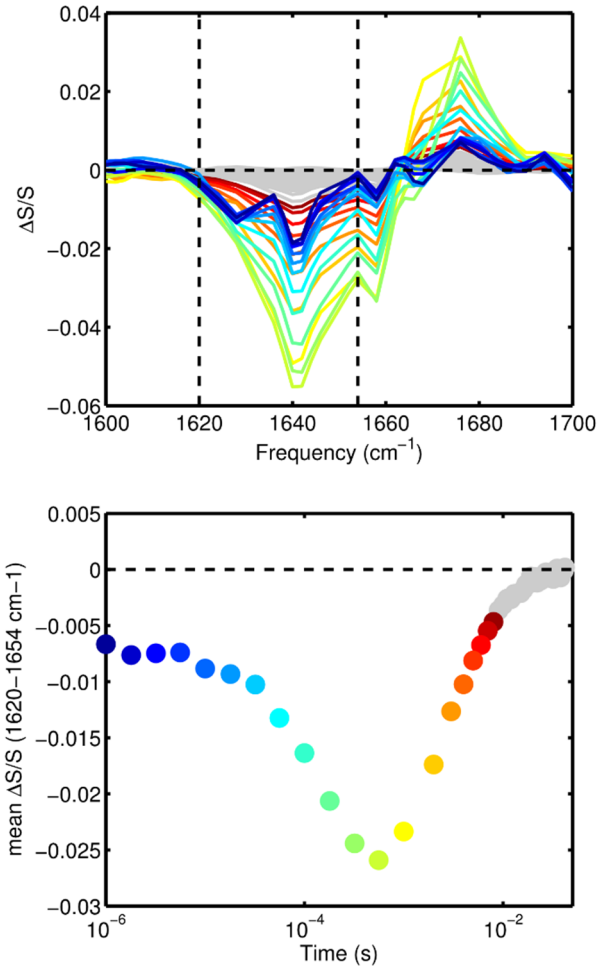

Figure S5: Temperature-jump HDVE spectral response of $NTL9_{1-39}$ at Ti=60 ºC along with the mean temporal response in the low-frequency region (1620-1654 cm^{-1}). The temporal response along with the exponential fit is also plotted in Figure 3 (main text).

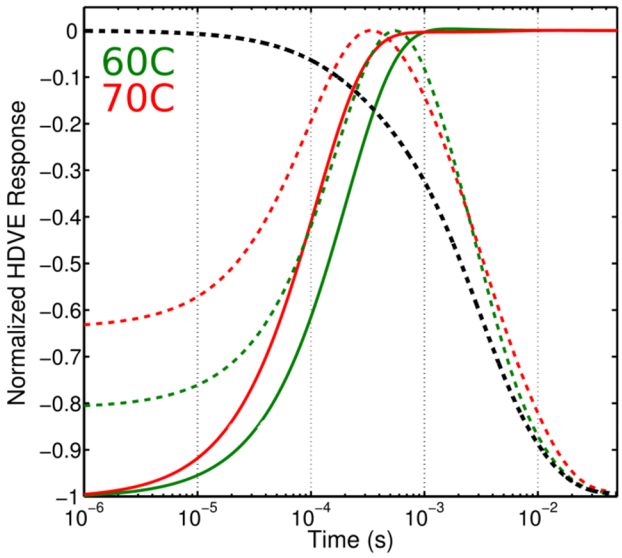

Figure S6: Temperature-jump HDVE temporal response of $NTL9_{1-39}$ at Ti=60 °C (green curves) and Ti=70 °C (red curves). The curves correspond to the average response between 1620 and 1654 cm^{-1}. The fitted water re-equilibration curve at 60 °C is shown by the black curve. Dashed curves represent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

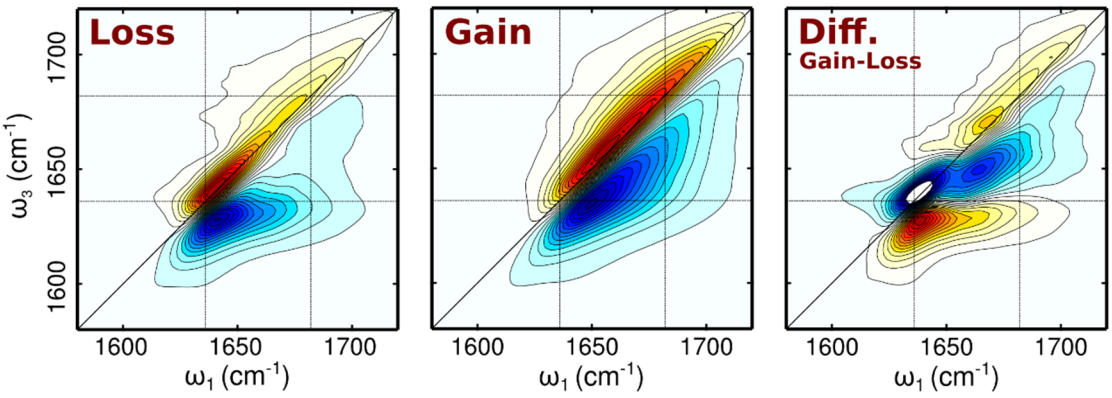

Figure S7: Spectra simulated using a reduced number of Markov states (1 folded, 10 disordered). See Figure 6 in the main text.

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.