Supporting Information for: Parallel folding pathways of Fip35 WW domain explained by infrared spectra and their computer simulation

Laura Zanetti-Polzi*[1], Caitlin M. Davis[2], Martin Gruebele[2,3], R. Brian Dyer[4], Andrea Amadei[5], and Isabella Daidone[1]

[1] *Department of Physical and Chemical Sciences, University of L'Aquila, via Vetoio (Coppito 1), 67010, L'Aquila, Italy*
[2] *Department of Chemistry and Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois*
[3] *Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois*
[4] *Department of Chemistry, Emory University, Atlanta, GA*
[5] *Department of Chemical and Technological Sciences, University of Rome "Tor Vergata", Rome, Italy*

Theory and Methods

The Perturbed Matrix Method to calculate IR spectra

The methodology used here to reconstruct amide I′ infrared spectra has been explained in details in previous articles.[1–3] Hereafter, the theoretical basis of PMM calculations and the computational procedure used to obtain vibrational spectra of solvated peptides are briefly outlined.

The MD-PMM approach is based on the combined use of quantum mechanical first principles and an extended phase space sampling provided by MD simulations. In PMM calculations, similarly to other mixed quantum-classical procedures,[4–6] a portion of the system is treated at the electronic level, the quantum center (QC), while the rest of the system is described at

*laura.zanettipolzi@univaq.it

a classical atomistic level and exerts an electrostatic perturbation on the QC electronic states. An orthonormal set of unperturbed electronic Hamiltonian ($\widetilde{H}^0$) eigenfunctions (Φ_j^0) are initially evaluated on the QC structure of interest which is typically the ground state equilibrium geometry. Indicating with $\mathcal{V}$ and $\mathbf{E}$ the perturbing electric potential and field, respectively, exerted by the environment on the QC (typically obtained by the environment atomic charge distribution and evaluated in the QC center of mass) we may, then, construct for each QC-environment configuration (as generated by explicit solvent MD simulation) the QC perturbed electronic Hamiltonian matrix ($\widetilde{H}$) as follows:

$$\widetilde{H} \simeq \widetilde{H}^0 + \widetilde{I} q_T \mathcal{V} + \widetilde{Z}_1 + \Delta V \widetilde{I} \tag{1}$$

$$[\widetilde{Z}_1]_{j,j'} = -\mathbf{E} \cdot \langle \phi_j^0 | \hat{\boldsymbol{\mu}} | \phi_{j'}^0 \rangle \tag{2}$$

where q_T and $\hat{\boldsymbol{\mu}}$ are the QC total charge and dipole operator, respectively, ΔV approximates all the higher order terms as a simple short range potential, $\widetilde{I}$ is the identity matrix and the angled brackets indicate integration over the electronic coordinates. The diagonalization of $\widetilde{H}$ provides a set of eigenvectors and eigenvalues representing the QC perturbed electronic eigenstates and energies.

For the more specific purpose of calculating amide I′ infrared spectra, trans-N-methylamide (NMA) was chosen as the QC model for each peptide backbone unit. The mass-weighted Hessian eigenvectors of the isolated trans-NMA molecule, calculated quantum chemically, provide the unperturbed vibrational modes of each peptide backbone unit from which the amide I mode is selected. Along such an eigenvector, a set of atomic configurations of the trans-NMA model were generated and, for each of these structures, an orthonormal set of unperturbed electronic Hamiltonian eigenfunctions were initially evaluated (see Unperturbed quantum chemical calculations section). Then, for each MD frame and for each IR chromophore (i.e., each of the N backbone peptide units), the perturbed electronic ground state and corresponding energy is calculated using Eqs. 1 and 2 for each peptide group after having fitted trans-NMA onto the WW peptide backbone unit. For each MD frame and for each IR chromophore the electronic ground state energy as a function of the mode coordinate is calculated, providing a perturbed energy curve that is modeled by a Morse potential thus yielding the vibrational frequency of the perturbed mode. Note that the side chain of the considered peptide group, the N-1 residues and the solvent define the

perturbing environment at each configuration generated by the MD simulation. Different secondary structure arrangements and/or different hydrogen bonding networks provide different perturbing environments to the considered peptide group, leading to a perturbed energy curve that is sensitive to the instantaneous conformation of the environment. Also the hydrogen bonds between the QC and its molecular environment are taken into account only by including the semiclassical interactions due to the atomic charges, i.e., no chemical bonding effects are considered.

To include the modes coupling effects due to interacting vibrational centers (i.e., excitonic effects) in the calculations, the evaluated perturbed frequencies are used to construct at each MD frame the excitonic coupling matrix describing the coupling among the QC modes by means of the transition dipole coupling (TDC) approximation, i.e., the expansion of the chromophore-chromophore interaction operator up to the dipolar terms. The perturbed frequencies for each oscillator, k, are used to include the excitonic effect by the construction and diagonalization of the excitonic Hamiltonian matrix (i.e., the Hamiltonian matrix for the interacting chromophores as expressed within the basis set provided by the products of the single chromophore perturbed vibrational eigenstates):

$$\widetilde{H} = \widetilde{I}\mathcal{U}_{vb,0} + \Delta\widetilde{H} \tag{3}$$

with $\mathcal{U}_{vb,0}$ the (vibronic) ground state energy of the interacting chromophores and $\Delta\widetilde{H}$ the excitation matrix whose diagonal elements are:

$$\left[\Delta\widetilde{H}\right]_{kl,kl} = h\nu_{kl} \tag{4}$$

and whose non-zero off-diagonal elements are given by the corresponding off-diagonal elements of the matrix representing the chromophores interaction operator (within the TDC approximation):

$$\hat{V}_{k,k'(k\neq k')} = \frac{\hat{\boldsymbol{\mu}}_k \cdot \hat{\boldsymbol{\mu}}_{k'}}{R^3_{k,k'}} - 3\frac{(\hat{\boldsymbol{\mu}}_k \cdot \boldsymbol{R}_{k,k'})(\hat{\boldsymbol{\mu}}_{k'} \cdot \boldsymbol{R}_{k,k'})}{R^5_{k,k'}} \tag{5}$$

as expressed in the same basis set of the excitonic Hamiltonian matrix, thus providing for each of such elements an interaction potential constructed by means of the single chromophore transition dipole moments. In Eq. 5 ν_{kl} is the kth chromophore lth excitation frequency, $\hat{\boldsymbol{\mu}}_k$ the kth chromophore dipole operator and $R_{k,k'}$ is the k' to k chromophore displacement vector defined

by the corresponding chromophores origins. In the excitonic Hamiltonian matrix only the first vibrational excitation of the electronic ground state for each chromophore must be involved, as higher vibrational excitations are forbidden and the coupling with excited electronic states may be neglected.

Diagonalization of the excitonic coupling matrix provides the instantaneous vibrational eigenstates and eigenvalues (now including vibrational mode coupling), and yields the perturbed vibrational frequencies and corresponding transition dipoles of the whole peptide. Note that this procedure allows to model the Hamiltonian eigenstates of the complex system including all the interacting chromophores using only quantum chemical calculations performed on the single amide groups.

Finally, the obtained perturbed excitation frequencies and corresponding transition dipoles are used to reconstruct the complete vibrational spectrum. Once the perturbed frequencies and transition dipoles are obtained at each MD frame, their distribution can be indeed evaluated using an appropriate number of bins in the frequency space providing the vibrational spectrum. Thus, the band width and line shape of the calculated spectra are obtained from the distribution of the perturbed frequencies as calculated via the MD-PMM approach at each MD frame and for each peptide group, avoiding the use of any empirical or adjustable parameter.

The main approximations of the above approach are the following. (a) The invariant mode approximation,[3] based on the assumption that the perturbations acting on a single IR chromophore do not significantly modify the forms of the vibrational modes (i.e., the eigenvectors of the mass-weighted Hessian) of interest, but rather alter only the corresponding frequencies. (b) The TDC approximation for calculating the excitonic coupling. Such an approximation is used also in other analogous calculations.[7–9] However, in most of the available TDC methodologies the excitonic coupling matrix is commonly constructed by using unperturbed single-residue vibrational states (i.e., in the absence of the environment perturbation), and the inclusion of the perturbation effects then typically involves different levels of phenomenological approximation, trying to optimize the computed–experimental matching.[7–9] In contrast, our excitonic coupling matrix is constructed from the basis set of the actual perturbed vibrational states, thus explicitly including the perturbation of the atomic-molecular environment in the definition of the basis set used to provide the excitonic states. (c) The vibrational coupling between chromophore and solvent modes is neglected, and therefore, the method does not properly treat chromophore vibrational modes involving

atomic coordinates of the first solvation shell.

Unperturbed quantum chemical calculations

The details of the unperturbed quantum chemical calculations were previously described[1] and are briefly summarized hereafter. As a model of the peptide group, i.e., the quantum center to be explicitly treated at electronic level, trans-NMA was chosen. Quantum chemical calculations were carried out on the isolated trans-NMA molecule at the Time Dependent Density Functional Theory (TDDFT) with the 6-31+G(d) basis set. This level of theory was selected because it represents a good compromise between computational costs and accuracy. The mass-weighted Hessian matrix was calculated on the optimized geometry at the B3LYP/6-31+G(d) level of theory and subsequently diagonalized for obtaining the unperturbed eigenvectors and related eigenvalues. The eigenvector corresponding in vacuo to the amide I′ mode was, then, used to generate a grid of points (i.e., configurations) as follows: a step of 0.05 a.u. was adopted and the number of points was set to span an energy range of 20 KJ/mol (in the present case 31 points). For each point, six unperturbed electronic states were then evaluated at the same level of theory providing the basis set for the PMM calculations.

Molecular dynamics simulations

The 100 μs-long MD simulation used in the present work was performed by the D. E. Shaw Research group[10;11] on the special-purpose machine Anton.[12] Fip35 was solvated in a cubic box with ≈50 Å side length containing ≈4000 TIP3P[13] water molecules and three chlorine ions to achieve a ≈30 mM ionic concentration. The simulations were performed using the Amber ff99SB-ILDN force field, which is based on the ff99SB force field.[14] All bonds involving hydrogen atoms were constrained to their equilibrium lengths with the SHAKE algorithm[15]. A cutoff of 9.5 Å for the Lennard-Jones and the short-range electrostatic interactions was used; for the long-range electrostatic interactions the k-Gaussian Split Ewald method was used[16]. The simulations were carried out in the NVT ensemble using the Nose-Hoover thermostat with a relaxation time of 1.0 ps saving frames every 200 ps. More details on the MD simulations can be found in the Supporting Information of the original work.[11]

Additional structural analyses

Helical population of the misfolded state

The DSSP analysis of the structures that populate the misfolded state reveals a relevant increase in helical conformations. The fraction of helical structures in the misfolded state (reported in Figure 4 of the main text) is indeed ≈27%. In the partially folded and unfolded states the same fraction is ≈13% and ≈5%, respectively, and negligible in the folded, H1F and H2F states.

The helical structures of the misfolded states have been thus more thoroughly characterized with the DSSP program. Such an analysis shows that both α-helical and 3-10 helical conformations are present in the misfolded state. The number of residues in each helical conformation is reported in Figure 1, A and B, as a function of time (considering only the frames that belong to the misfolded state). In Figure 1, C and D, the corresponding number of helical turns is also reported (the average number of residues per helical turn is 3.6 for α-helical conformations and 3 for 3-10 helical conformations). The average number of residues in α-helical conformations is 7.15, corresponding to an average number of turns equal to 1.99 with at least one alpha-helix turn in 83% of the frames belonging to the misfolded state. The average number of residues in 3-10 helical conformation is 4.9, corresponding to an average number of turns equal to 1.63 with at least one alpha-helix turn in 67% of the frames belonging to the misfolded state.

Analysis of the population of the H1F and H2F states

As mentioned in the main text, the normalized distribution of the RMSD of the C_α atoms of the three-stranded core residues with respect to the folded structure (reported for both H1F and H2F states in Figure 5A of the main text) shows for both states the presence of a low RMSD and a high RMSD peak.

The low RMSD peak is very similar for the two states and is located at ≈0.3 nm (i.e., with a small deviation with respect to the folded state three-stranded core). The structure of the highest populated cluster for the low RMSD ensemble, as obtained by applying the RMSD-based clustering procedure included in the GROMACS package,[17] is reported in Figure 2 for both H1F (A, red) and H2F (B, blue) together with the comparison with a representative structure of the folded state (in gray) and both show the

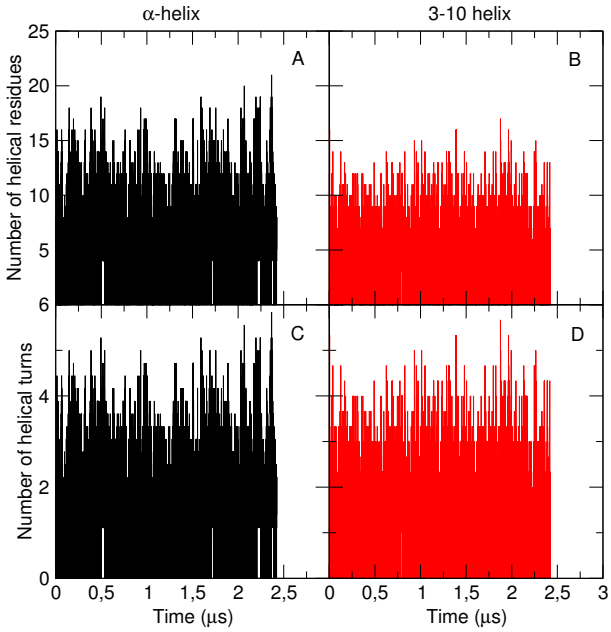

Figure 1: Number of residues (A and B) and turns (C and D) in α-helical (A and C) and in 3-10 helical (B and D) conformation in the frames belonging to the misfolded state as a function of time.

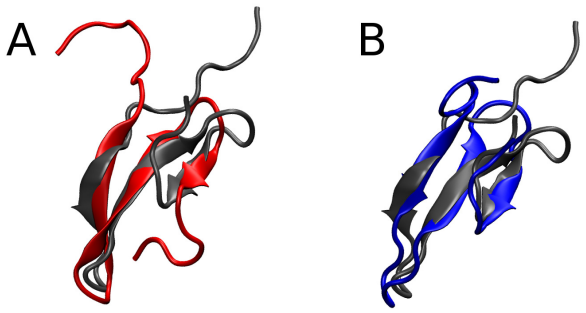

Figure 2: Structure of the highest populated cluster as obtained from clustering analysis of the structures belonging to the low RMSD peak (RMSD below 0.36) of H1F (A, red) and H2F (B, blue). The two structures are superimposed to a representative structure of the folded state (in gray).

formation of the three-stranded core.

The high RMSD peak is different in the two states. The H1F state shows a very broad distribution centered at ≈0.6 nm while the H2F state shows a sharper peak centered at ≈0.45 nm (see Figure 5A of the main text). The analysis of the spectra calculated for both H1F and H2F on the high RMSD conformations and the analysis of the secondary structure content of the two states (see Figure 4 of the main text) suggest that such a difference is due to the presence of residual β structure in hairpin 1 in the H2F state while, on the contrary, in the H1F state hairpin 2 is on average more unstructured.

This has been verified by monitoring with the DSSP program the number of residues with β-sheet or β-turn conformation in hairpin 2(hairpin 1) in the H1F(H2F) state with high RMSD with respect to the folded state. Such an analysis shows that for H1F with high RMSD 45% of the MD frames have at least 2 β-sheet or β-turn residues in hairpin 2 (partial formation of the three-stranded core) and only 9% of the MD frames have the entire core formed (4 or more residues in β-sheet or β-turn in hairpin 2). For H2F a consistent raise in the same populations can be observed: 63% of the MD frames have at least 2 β-sheet or β-turn residues in hairpin 1 (partial formation of the three stranded core) and 28% of the MD frames have the entire core formed (4 or more residues in β-sheet or β-turn in hairpin 1). The residues that are monitored for the above analysis are residue 7 to 12 for hairpin 1 and 26 to 31 for hairpin 2. As previously mentioned, the folded state three-stranded core residues are 8 to 11, 19 to 22 and 27 to 30. One additional

residue at each side of the three-stranded core fragment for both hairpin 1 and hairpin 2 has been considered in the analysis in order to include also possible conformations in which a non-native three-stranded core is formed (e.g., misregistered contacts).

References

[1] I. Daidone, M. Aschi, L. Zanetti-Polzi, A. Di Nola, and A. Amadei. On the origin of IR spectral changes upon protein folding. *Chem. Phys. Lett.*, 488:213–218, 2010.

[2] A. Amadei, I. Daidone, A. Di Nola, and M. Aschi. Theoretical-computational modelling of infrared spectra in peptides and proteins: a new frontier for combined theoretical-experimental investigations. *Curr. Opin. Struct. Biol.*, 20:155–161, 2010.

[3] A. Amadei, I. Daidone, L. Zanetti-Polzi, and M. Aschi. Modeling quantum vibrational excitations in condensed-phase molecular systems. *Theor. Chem. Acc.*, 129:31–43, 2011.

[4] J. Gao and D. G. Truhlar. Quantum mechanical methods for enzyme kinetics. *Ann. Rev. Phys. Chem.*, 53:467–505, 2002.

[5] T. Vreven and K. Morokuma. Chapter 3 hybrid methods: ONIOM(QM:MM) and QM/MM. *Ann. Rep. Comp. Chem.*, 2:35–51, 2006.

[6] H. M. Senn and W. Thiel. QM/MM studies of enzymes. *Curr. Opin. Struct. Biol.*, 11:182–187, 2007.

[7] Christopher M. Cheatum, Andrei Tokmakoff, and Jasper Knoester. Signatures of β-sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J. Chem. Phys.*, 120(17):8201–8215, 2004.

[8] Chewook Lee and Minhaeng Cho. Local amide i mode frequencies and coupling constants in multiple-stranded antiparallel β-sheet polypeptides. *J. Phys. Chem. B*, 108(52):20397–20407, 2004.

[9] Eeva-Liisa Karjalainen, Harish Kumar Ravi, and Andreas Barth. Simulation of the amide i absorption of stacked β-sheets. *The Journal of Physical Chemistry B*, 115(4):749–757, 2010.

[10] Stefano Piana, Krishnarjun Sarkar, Kresten Lindorff-Larsen, Minghao Guo, Martin Gruebele, and David E. Shaw. Computational design and experimental testing of the fastest-folding β-sheet protein. *J. Mol. Biol.*, 405(1):43–48, 2011.

[11] David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[12] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–11, 2009.

[13] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[14] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.*, 65(3):712–725, 2006.

[15] Ross A Lippert, Kevin J Bowers, Ron O Dror, Michael P Eastwood, Brent A Gregersen, John L Klepeis, Istvan Kolossvary, and David E Shaw. A common, avoidable source of error in molecular dynamics integrators. *J. Chem. Phys.*, 126(4):046101, 2007.

[16] Yibing Shan, John L Klepeis, Michael P Eastwood, Ron O Dror, and David E Shaw. Gaussian split ewald: A fast ewald mesh method for molecular simulation. *J. Chem. Phys.*, 122(5):054101, 2005.

[17] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F. van Gunsteren, and Alan E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.*, 38(1-2):236–240, 1999.