The Atomic Detail of Protein Folding Revealed by an Ab Initio Reappraisal of Circular Dichroism

Alan Ianeselli,[†] Simone Orioli,[‡,§] Giovanni Spagnolli,[†] Pietro Faccioli,[*,‡,§] Lorenzo Cupellini,[¶] Sandro Jurinovich,[¶] and Benedetta Mennucci[*,¶]

†*Centre for Integrative Biology, Trento University, Via Sommarive 9, 38128 Povo (Trento)*
‡*Physics Department, Trento University, Via Sommarive 14, 38128 Povo (Trento)*
¶*Dipartimento di Chimica e Chimica Industriale, University of Pisa, via G. Moruzzi 13, 56124, Pisa, Italy*
§*INFN-TIFPA, Via Sommarive 14, 38128 Povo (Trento)*

E-mail: pietro.faccioli@unitn.it; benedetta.mennucci@unipi.it

Abstract

Circular Dichroism (CD) is known to be an excellent tool for the determination of protein secondary structure due to fingerprint signatures of α and β domains. However, CD spectra are also sensitive to the 3D arrangement of the chain as a result of the excitonic nature of additional signals due to the aromatic residues. This double sensitivity, when extended to time-resolved experiments, should allow protein folding to be monitored with high spatial resolution. To date, the exploitation of this very appealing idea has been limited, due to the difficulty in relating the observed spectral evolution to specific configurations of the chain. Here, we demonstrate that the combination of atomistic molecular dynamics simulations of the folding pathways with a quantum chemical evaluation of the excitonic spectra provides the missing key. This is exemplified for the folding of canine milk lysozyme protein.

1 Introduction

More than 50 years after protein folding was proven to be a spontaneous process[1], a general agreement on the underlying molecular mechanisms has not been reached yet[2–4]. In the quest to solve this uncertainty and achieve a complete understanding at the required spatiotemporal resolution, a central role is played by the combination of atomistic computer simulations and experimental measurements.[5] Unfortunately, such a combination is still quite challenging, due to both computational and experimental limitations.

On the one hand, the complexity combined to the large dimension of proteins hampers the application of accurate quantum-mechanical approaches to the study of their energy surfaces and dynamics, forcing the use of classical descriptions based on molecular mechanics force fields. Moreover, even adopting such simplified models, computer simulations can only cover relatively short time intervals, up to ms.[6,7] Consequently, additional approximations need to be introduced in order to bridge the gap between the biologically relevant and the computationally accessible time scales (see e.g. Refs. [8–18] and references therein). All the limitations intrinsic to these approximated methods have so far prevented any of them to become consensually accepted.

On the other hand, the available experimental techniques either have a low spatial resolution or can probe with high resolution only point-to-point distances. For example, hydrogen-deuterium exchange detected by mass spectroscopy[19] provides information about the solvent accessible regions, while small-angle scattering combined with the stopped-flow technique measures the overall degree of compactness of the polypeptide chain.[20] However, these methods lack the resolution required to thoroughly assess the predictions of atomistic models. Furthermore, the limited time-resolution restricts their applicability to relatively slow structural reactions. A few alternative experimental techniques have been developed to probe specific distances, with much higher spatial and temporal resolution. In particular, single-molecule Förster Resonance Energy Transfer (smFRET) experiments can measure sub-nanometric variations of the distance between two chromophores located at specific positions along the chain, with a time resolution in the μs scale.[21,22] Atomic force microscopy can also measure with sub-nanometric resolution the relative distance between two residues, subject to an externally applied mechanical stress.[23,24] In general, the development of these single-molecule techniques have brought inestimable new insight into protein folding kinetics and thermodynamics.[25,26] However, they require to alter the chemical structure of the polypeptide chain, by implementing point mutations, and attaching fluorescent probes or molecular "handles". The question then arises whether such chemical manipulations can significantly affect either the dynamics or the biology of the protein under investigation.

In contrast, direct spectroscopic measurements do not require any chemical transformation. Circular dichroism (CD) experiments are routinely performed to probe the amount

of secondary structure elements in equilibrium ensembles of proteins.[27,28] Namely, using empirical algorithms it is possible to estimate the average content of α-helices and β-sheets, from the CD spectrum in the far-UV region (190–230 nm)[29–31]. Moreover, structure-based calculations are able to predict CD signals that arise from secondary structure, particularly for helical proteins.[32–34] Similarly, time-resolved CD experiments can be used to probe the kinetics of secondary structure formation, in protein folding reactions which are sufficiently slow to be monitored by the stopped flow technique.[35–39] Unfortunately, this kind of analysis yields only partial information about the folding mechanism, as it only "sees" the secondary structure, which, in most cases, forms at the early stages of the protein folding reaction. As a matter of fact, CD is also sensitive to the tertiary structure; in particular, the signal due to the excitonic coupling of the π-π^* excitations on the aromatic rings of phenylalanine, tyrosine and tryptophan could represent an extremely responsive probe for small structural changes[40,41]. However, the signal due to the secondary structure is so dominant in the typical 190–230 nm region that a clear disentanglement of the two sources is very difficult. In principle, one could resort to a different spectral region such as the 240–300 nm window, where the CD is determined only by the aromatic residues. In any case, however, the pure analysis of the CD spectra cannot give an univocal description of the three-dimensional rearrangement of the chain, but one must rely on some *a priori* knowledge.[39]

Here we present a new way to look into the time resolved CD spectra by introducing a theoretical scheme which predicts the folding path and follows the evolution of the CD spectrum. The scheme is *ab initio* in the sense that all the required ingredients (the protein configurations and the optical spectra) are deduced directly from atomistic calculations. We start by generating an ensemble of protein folding trajectories, using a state-of-the-art all-atom force field. These calculation are made computationally feasible by adopting a recently developed biased molecular dynamics based on a variational approximation.[42,43] Then, we analyze the resulting trajectories to identify long-lived intermediates and collect representative configurations. The structures thus obtained are used to simulate the CD spectra of long-lived states with an exciton model based on quantum chemical calculations.[44–46]

We test our approach on the folding of canine milk lysozyme, a 129 residue globular protein with folding time in the time scale of seconds.[47,48]

Remarkably, our *ab initio* calculations reproduce all the essential features observed in the time resolved CD measurements, showing that the CD signals due to the excitonic coupling of the aromatic residues can indeed be used to reveal extremely small changes in the tertiary structure of the chain. This clearly demonstrates that experimental time-resolved CD spectra when combined with advanced path sampling and quantum electronic structure techniques provide an extremely powerful approach for the determination of protein folding pathways, with a spatial resolution higher than the fraction of nm.

2 Methods

2.1 Bias Functional Approach

The Bias Functional (BF) approach is based on combining a special kind of biased Molecular Dynamics (MD) called Ratchet-and-Pawl MD (rMD) with a variational principle rigorously derived from the path integral representation of Langevin dynamics. Within this framework, the calculation of protein folding trajectories is performed according to a three-step procedure (see also Section S1 of the Supporting Information). First, an ensemble of unfolded configurations is obtained by thermal unfolding MD simulations, initiated from the energy-minimized native structure. Then, for each of such denatured configurations an ensemble of *trial* folding pathways is generated by means of rMD simulations.[49,50] In the rMD, the chain is left free to evolve according to the unbiased force field any time it spontaneously increases the struc-

tural overlap with the native state, defined in terms of a suitable collective variable, closely related to the fraction of native contacts. Conversely, a history dependent harmonic bias is introduced to discourage the diffusion towards configurations with lower native overlap. Finally, all the rMD trajectories generated from each given unfolded configuration are scored by evaluating their Bias Functional, defined as

$$T = \sum_{i=1}^{N} \frac{1}{m_i \gamma_i} \int_0^t d\tau |\mathbf{F}^i_{rMD}(X,\tau)|^2. \quad (1)$$

In this equation, γ_i and m_i are respectively the viscosity and mass of the i-th particle, while $|\mathbf{F}^i_{rMD}(X,\tau)|^2$ is the square-modulus of the rMD force, integrated for all times along the rMD trajectory $X(\tau)$. It can be shown that the rMD trial trajectories with the lower value of this functional are those with the higher probability to occur in completely unbiased simulations. In this sense, such Least Biased Trajectories (LBT) provide a variational estimate of the folding pathways generated by the Langevin dynamics. The accuracy of the BF approach has been assessed against the results of plain MD simulations performed on the Anton supercomputer[42,43] and against experimental data concerning different protein folding reactions[51,52] and other conformational transitions.[52]

All simulations were performed using the Amber ff99SB-ILDN force field[53] with the implicit solvent model implemented in GROMACS 4.6.5.[54] In such an approach, the Born radii are calculated according to the Onufriev-Bashford-Case algorithm.[55] The hydrophobic tendency of non-polar residues is taken into account through an interaction term proportional to the atomic solvent accessible surface area. The solvent exposed surface of the different atoms is calculated from the Born radii, according to the approximation developed by Schaefer, Bartels, and Karplus in Ref. 56. We implemented the collective variable used in the BF approach within PLUMED 2.0.2.[57]

2.2 Quantum chemical calculations

The CD spectrum of each selected configuration is calculated employing an excitonic model. The exciton Hamiltonian is written on the basis of the π-π^*excitations of the aromatic residues (phenylalanine, tyrosine and tryptophan). Exciton couplings between excitations belonging to different residues are calculated as Coulomb interactions of the corresponding transition densities.[58,59] The calculated electric transition dipole moments of the local excitations are then combined with the exciton coefficients to obtain the CD intensities as detailed in Section S2 of the Supporting Information. Local excitation properties and exciton couplings were computed in the time-dependent Density Functional Theory (TDDFT) with the CAM-B3LYP[60] functional and the 6-31+G(d) basis set. The Integral Equation Formalism Polarizable Continuum Model (IEF-PCM)[61] was used to describe the effect of the environment on both site properties and couplings, assuming the dielectric properties of water.

The aromatic side-chains of Trp, Tyr, and Phe were modeled respectively as 3-methylindole (3MI), p-cresol, and toluene. Model structures were optimized with the B3LYP functional[62] and the cc-pVDZ basis set. Before calculating site properties and couplings, we projected the optimized model structures on the aromatic side-chain geometries from the MD simulation according to the RMSD criterion. In our excitonic model, we include the four π-π^* transitions of each aromatic side chain that occur in the near- and far- UV region. These are assigned respectively to two low-energy states L_b and L_a, and two high-energy states, B_b and B_a.[63] As the most common DFT functionals gives an inhomogeneous description of the different π-π^*excitation energies (including a wrong ordering of the two L states[64]) the values of the site energies have been fitted to the experimental values (see Section S2 of the SI). All quantum chemical calculations were performed using Gaussian16[65] whereas the excitonic spectra were generated with the EXcitonic Analysis Tool (EXAT).[66]

3 Results

For clarity, the presentation of the results is divided into two sections, one referring to the folding trajectories, for which the Bias Functional Approach was used used, and the other focused on the CD spectra, for which the excitonic model was employed. Finally, an additional analysis on an "artificial" protein is presented to quantify the sensitivity of the proposed strategy to specific characteristics of the folding mechanism.

3.1 Folding trajectories and metastable states

The canine milk lysozyme is composed of an α-domain with four α-helices, and a β-domain formed by three antiparallel β-strands: the crystal structure of this protein is shown in Figure 1. Experimental CD studies of kinetic refolding on this protein reported in Ref. 47 revealed the existence of different kinetic relaxation time scales; a very fast initial phase associated to the formation of a so-called burst intermediate (**I-Burst**) – which may be identified with the protein's molten globule state – and a later folding intermediate (**I-Second**), populated at a rate of ~ 22 s^{-1}. Finally, the system relaxes to the **Native** state, at a rate of ~ 0.5 s^{-1}. Kinetic analysis allowed to pinpoint the difference between the CD spectra of **Native** and **I-Second**, which was related to changes in the tertiary structure, while the secondary structure remains unchanged. The spectral changes observed during this last refolding step were attributed to the coupling among the tryptophan residues of the β-domain (See Figure 1)[47].

In order to simulate the folding process and identify the metastable states, a two-step protocol is used. As a first step, we generated 10 independent unfolded configurations for the protein, as described in the Methods Section. For each of them, we used the rMD to generate many trial folding trajectories. Then, we applied the BF variational principle to identify the rMD folding pathway with the largest probability to be realized in the *unbiased* dynamics. In the second step, the generated trajectories were analyzed in order to identify all metastable states that are expected to be sufficiently long-lived to contribute to the CD spectrum.

In Figure 2 we report the probability density $P(Q, R)$ on the plane identified by the fraction of native contacts Q and the RMSD to the crystal native structure R. This was calculated using all configurations in all the rMD trial trajectories generated during the BF calculation. Even if this density plot is not directly related to the thermodynamic potential of mean-force (the trial trajectories do not represent equilibrium transitions), it can still be used as a qualitative indicator of the existence of long-lived metastable states.

The basin at the top left corner of the plot, which we refer to as the **Unfolded** state, is populated by fully denatured configurations, which are reached only by high-temperature MD trajectories initiated from the native state. Arguably, this state is not visited by spontaneous unfolding/refolding trajectories at room temperature. On the other hand, it may be regarded as a model of the denatured state which is experimentally reached by temperature jump or chemical denaturation techniques.

A first metastable state along the folding pathway (which we identify with the **I-Burst** state), contains an ensemble of relatively compact configurations. This state is characterized by a high structural heterogeneity, with a significant content of secondary structure, as shown by the randomly selected conformers reported in Figure 2. The formation of native secondary structures is completed in a second intermediate state, which we identify with the **I-Second** state. Finally, in order to reach the native state, the chain needs to establish a number of tertiary contacts, while the amount of secondary structural elements remains essentially unaltered.

To highlight the key structural differences between the **I-Second** and the **Native** state, we randomly harvested about 100 configurations from the LBTs obtained after applying the variational principle using the criterium detailed in the Supporting Information. These configurations are shown as dots in the density plot and the corresponding three dimensional structures

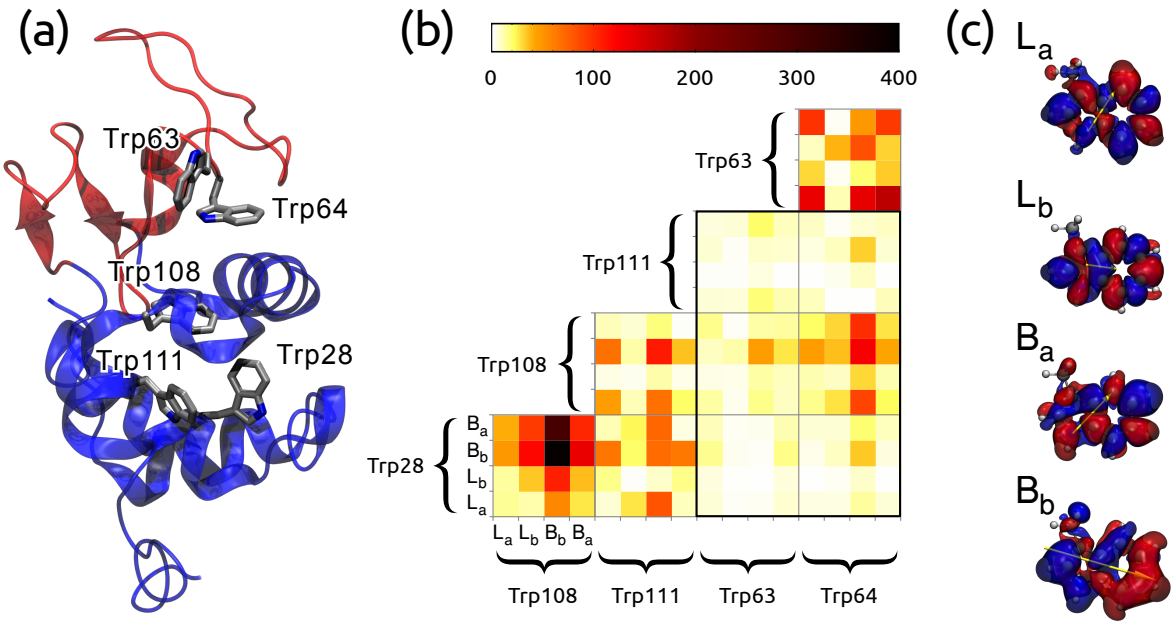

Figure 1: (a) Crystal structure of the canine milk lysozyme (PDB code: 1EL1):[67] the α and β secondary structure are highlighted in blue and red, respectively, whereas the tryptophan residues are depicted in tube representation. (b) Map of the the electronic couplings (cm^{-1}) among the four π-π^*excitations (L_a, L_b, B_a and B_b) of the five different tryptophan residues (Trp28, Trp108, Trp111, Trp63, and Trp64). (c) Graphical representation of the transition densities corresponding to the L_a, L_b, B_a and B_b excitations. The arrows denote the associated transition dipoles.

are shown in light grey in Figure 2. In Figure 3 we report the difference between the average contact maps in these two states, which reveals the relevant tertiary contacts formed during the **I-Second→Native** transition. These results indicate that the native state is reached by the tertiary packing of α- and β-domains, consistently with the conclusions reported in Ref. 47 from the analysis of time-resolved CD spectra.

3.2 Excitonic CD spectra

Once the main long-lived states have been revealed, we generate the corresponding CD spectra, which can directly be compared to experiments. CD spectra are obtained by performing quantum chemical calculations on the set of configurations generated as described above. To this goal, we applied the excitonic model detailed in the Methods and the Section S2 of the Supporting Information. This approach was successfully applied to predict optical spectra of biological polymers[45,68] and light-harvesting pigment-protein complexes[46,69,70], in ordinary equilibrium conditions.

Here the chromophoric units are the aromatic residues (phenylalanine, tyrosine and tryptophan), characterized by their four lowest π-π^*excitations. As we show in the Figure S7 of the Supporting Information, the tryptophan (Trp) residues are responsible for all of the main features of the CD spectra, while the other aromatic residues only slightly affect intensity redistribution among the different peaks. For this reason, in the following analysis, we will use the results obtained by only including the Trp residues. This simplification allows us to make a more direct assessment of the relation between the CD signal and the 3D packing structure of the protein.

The indole chromophore of tryptophan exhibits four readily identifiable π-π^* electronic transitions in the 200-300 nm region. According to the common notation for aromatic chromophores, these transitions are collected into two pairs called, L_a and L_b, and B_a and B_b, respectively. Transitions from the ground state

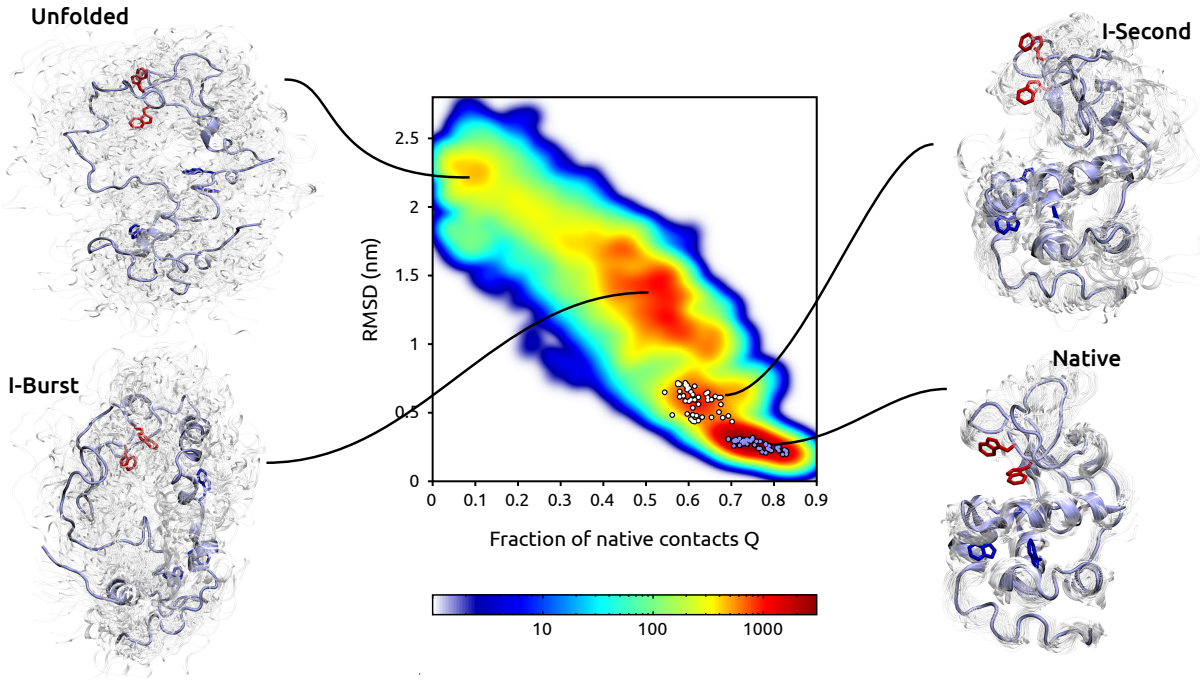

Figure 2: Density plot obtained by projecting the full ensemble of rMD trajectories onto the plane defined by the RMSD to the crystal native structure and the fraction of native contacts Q. The high density regions reveal the existence of long-lived intermediates, denominated **Unfolded**, **I-Burst**, and **I-Second** states, respectively. The structures at the side correspond to randomly selected representative configurations in each of such states.

to the L states are weakly allowed while the B states are accessible from the ground state by fully allowed transitions. In the right panel of Figure 1 we show the corresponding transition densities and the related transition dipoles.

In the experimental CD spectrum a weak positive band in the near-UV region appears, due to the excitonically coupled L_a and L_b states: this positive band is typical of lysozimes.[47,48,71,72] The B_a and B_b states contribute to the CD signal in the far-UV region, where it overlaps with the large protein signal coming from the exciton coupling between n-π^* and π-π^* transitions of amide chromophores.

We report in Figure 4(a) the CD spectrum due to the excitonic coupling of all the Trp residues calculated on the crystal structure. The values of the couplings for all the different pairs of transitions and the different Trp residues are shown in Figure 1.

The spectrum presents a weak positive band at $\sim$ 290 nm, which is in agreement with the characteristic band of lysozimes. A strong bisignate couplet (about one order of magnitude stronger than the $\sim$ 290 nm band) arising from the coupling between the B transition is found at shorter wavelengths. In the experiment, this couplet is covered by the strong negative signal due to the amide transitions.

To investigate the origin of the spectral shape, we decoupled the signals coming from the Trp residues in the α- domain (i.e. at location 28, 108 and 111, see Figure1) from those in the $\beta-$domain (i.e. at location 63 and 64). The spectra of these two contributions are shown by the colored lines in Figure 4(a)). They show almost opposite signs and are slightly shifted relative to each other. Notably, the sum of these two decoupled spectra (gray dotted line) significantly differs from the fully coupled spectrum, which includes the couplings between Trp in the α and β domains (black solid line). This difference clearly indicates that the coupling between the two domains (see also Figure 1) is not negligible in determining the CD signal, and confirms the sensitivity of the near UV signal to the specific three-dimensional re-arrangement of the $\alpha-$ and $\beta-$ domains. In particular, the

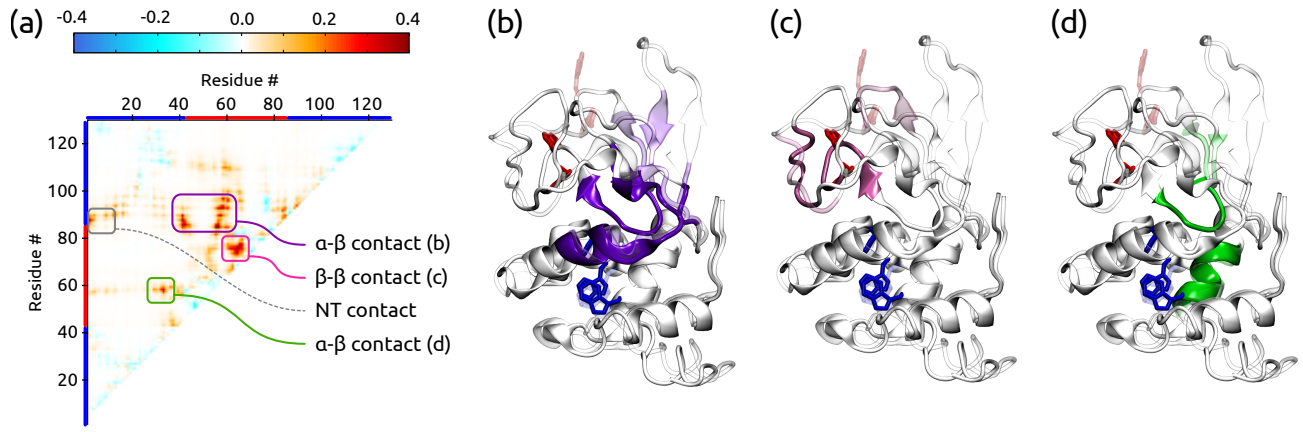

Figure 3: (a) Difference between the average α-carbon contact map in the **I-Second** and **Native** states, highlighting the key tertiary contacts between $\alpha-$ and $\beta-$ domain formed during the **I-Second**→**Native** transition. (b,c,d) Superimposed structures of **I-Second** (transparent) and **Native** (opaque) showing in color the residues involved in the tertiary contacts, with the same color code as (a). The Trp residues of both domains are shown, with the same color code as in Figure 1.

interactions between these two domains determine half of the strength of the $\sim$ 290 nm positive band.

An important issue concerns the statistical significance of the calculated CD spectra.[73,74] While the experimental signal corresponds to an ensemble average over a huge number of independent protein configurations, the calculated spectra discussed so far were obtained from a single (crystal) structure. To quantify the role of structural heterogeneity in the simulated spectra, in Figure 4(b) we report the CD spectrum obtained by averaging the signal calculated from the structures in the **Native** state, which were reached by independent folding trajectories and harvested using the criterion described above. The CD spectrum (black line) retains the overall shape of that calculated from the single crystal configuration, albeit with a general reduction of intensity. The decoupled spectra for the two domains also maintain the same shape over the whole wavelength range. The reduction of CD intensities is the result of structural disorder, which affects both couplings and relative orientations. We note that, given the very small intensity of the positive band at $\sim$ 290 nm in the experiments, the fluctuations within the **Native** structures give rise to a large variability in the signal over 260 nm, and tend to suppress the small positive signal. Looking into more details, we see that the contribution to this signal from the Trp residues in the α-domain is only marginally reduced compared to the crystal structure, whereas the contribution from those in the β-domain is suppressed by more than a factor of two. This indicates a larger structural disorder for the β-domain. Interestingly, the **Native** configurations selected to compute this spectrum look very similar by visual inspection (see right panel of Figure 2), demonstrating that the near-UV CD signal is indeed able to resolve very small structural differences.

In Figure 4(c) the average CD spectrum calculated from the configurations in the **I-Second** state is reported, along with the contributions from the two domains, as in the previous analysis. The CD spectrum of the α-domain Trp residues is virtually identical to that of the **Native** structures. On the contrary, the β-domain contribution is noticeably broadened and reduced in intensity, even though the general shape is maintained. Most importantly, the full spectrum of **I-Second** is significantly different from the **Native**, both in the near-UV region and in the region of the B-couplet around 200–230 nm.

As the sum of α-domain and β-domain spectra in the **I-Second** equals the fully coupled spectrum, the two domains must be excitoni-

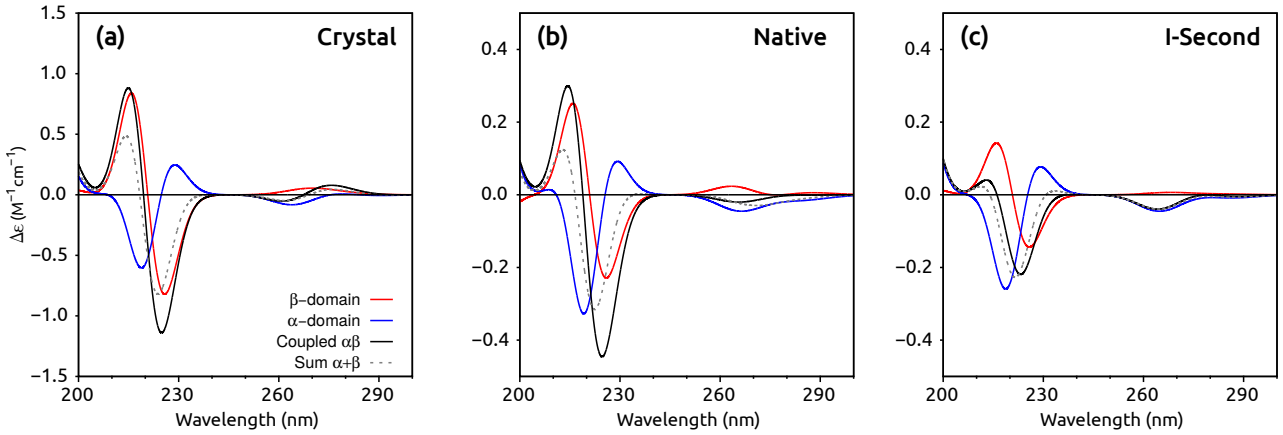

Figure 4: (a) Simulated CD spectra at the crystal structure; (b) average CD spectra from 50 **Native** structures identified by our variational calculation (i.e. using the LBTs) (c) and average CD spectra of 47 **I-Second** structures still obtained with the same variational calculation. Black lines refer to the spectrum including all the coupled Trps, blue lines refer to the α-domain (Trp 28–108–111), and red lines to the β-domain (Trp 63–64) Trps, gray dotted lines refer to the sum of the two decoupled domains.

cally decoupled in **I-Second**, in contrast with what is observed in the **Native** state. This feature can be understood by analysing the pairwise distances of the Trp residues in the two structures (see Figure S8 in the Supporting Information). Indeed, while the intra-domain Trp distances are almost the same in the **Native** and in the **I-Second**, the inter-domain distances are much larger in **I-Second** with respect to the **Native** structures.

For both **Native** and **I-Second** state, the convergence of the CD spectra have been checked by comparing averages obtained on an increasing number of independent configurations. This check (reported in Figure S9 in the SI) confirms a clear convergence for the **Native** state with the selected 50 configurations. As expected due to the larger flexibility, the convergence of the **I-Second** seems slower but, also in this case, the set of configurations used gives a robust description of the main features of the spectrum.

The results obtained from the CD analysis performed on the Trp suggest that in the **I-Second** state the native secondary structure is already reached as well as the tertiary structure within two independently folded sub-units (foldons). The main structural difference between **I-Second** and **Native** is the packing of the two foldons which is reached in the **Native** state, but not in **I-Second**. These results agree with the structural analysis performed by inspection of the folding trajectories.

In order to directly compare with the experiments[47] we computed the so-called kinetic difference spectrum, i.e. the difference between **Native** and **I-Second** spectra. The experimental kinetic difference is a bisignate couplet with a negative peak at ∼230 nm and a positive peak at ∼210 nm, with zero-crossing at 226 nm. Both α-Helix and β-strand secondary structures give negative contributions in the 210–240 nm range, therefore a difference in the secondary structure cannot give a bisignate difference between CD spectra with a zero-crossing at 226 nm. The kinetic difference is therefore due to the change in the packing of Trp residues in the tertiary structure.[47,71,75] On the basis of this observation, we have computed the kinetic difference by considering only the signal due to the Trps. The results are reported in Figure 6(a) together with the experimental data. The computed kinetic difference, despite a small ∼5 nm blue shift, compares well in shape and intensity with the experimental points, showing that the structure of the **I-Second** state is adequately captured by the configurations visited by the folding trajec-

tories generated with our variational approach.

3.3 Protein composition and sensitivity

To quantify the sensitivity of the proposed approach to specific characteristics of the folding process, we have repeated the same procedure for an artificial system: the four disulfide bridges (Cys30–Cys115, Cys127–Cys6, Cys94–Cys76 and Cys65–Cys80) present in the original protein were "reduced" and new unfolded configurations were computed by thermal unfolding. As for the original protein, we have first generated an ensemble of folding pathways, which was analyzed in order to identify the long-lived metastable states. A statistical analysis of these new folding trajectories is reported in Section S3 of the Supporting Information. A representative set of configurations sampled from these states has been finally used in order to calculate the corresponding CD spectra.

Coherently with the simulations in presence of disulfide bridges, four long-lived states can be identified (see Fig. 5a): an **Unfolded** one, around $Q \sim 0.1$ and RMSD ~ 4 nm, a very extended molten globule state centered in $Q \sim 0.5$ and RMSD $\sim$ 2 nm (the **I-Burst** state), another folding intermediate around $Q \sim 0.7$ and RMSD ~ 1 nm (the **I-Second** state) and a **Native** configuration. A more detailed structural analysis, however, unveils a rather different folding mechanism. Indeed, the folding reaction now proceeds by stabilizing in sequence the β and α regions, in the inverted order with respect to the previous case. This new folding pathway is made possible by the lack of a constraint between the CYS127 and CYS6 positions, which makes the C-terminal region much more mobile. As a result, this part of the chain becomes the last to form tertiary contacts. These differences in the mechanism lead to a completely different three-dimensional arrangement of the **I-Second** intermediate, which is depicted in Fig. 5c : the C-terminal region is not connected to main body of the protein by a disulfide bond and it is therefore much more flexible.

The differences in the structure and the flexibility of the metastable states are reflected in the kinetic difference of CD signals, as shown in Figure 6(b).

These results suggest that, in the absence of disulphide bonds, **I-Second** state displays a more native-like structure. This is confirmed by inspection of an ensemble of configurations in the **I-Second** (see Fig. 5c), which shows that the two foldons are already packed in the **I-Second** state.

As a further validation of the proposed approach we have analyzed a different protein, the Im7 belonging to the family of colicin immunity binding proteins of *Escherichia coli*. Im7 provides a good additional test, because, contrary to lysozyme, it contains only one Trp residue and the CD spectrum in the aromatic region (240–320 nm) is determined by the interaction of this Trp with other aromatic residues. The folding of Im7 was previously studied by some of the present authors[51] through the same Bias Functional approach used for the natural and the artificial lysozime proteins. The results of our calculations and the comparison with experimental data are reported in Section S4 of the SI. Here we only summarize the main findings. To highlight the CD changes following the Intermediate $\rightarrow$ Native transition, we have calculated the difference CD spectrum in analogy to what reported in 4(a)-6(a) for the two lysozimes. Once more, the calculations correctly reproduce the experimental CD difference,[76] despite a small blue shift, which can be imputed to the lacking of vibronic effects. For the very weak CD signals in this region of the spectrum, these effects are much more visible than in the region investigated for lysozine.[77] The analysis of the origin of the CD difference has also allowed us to confirm the interpretation given in Ref. 76 based on experiments on the mechanism of folding. In particular, we found that among the four native α-helices, only helices I and II are found to be folded and associated in a highly native like manner in the long-lived intermediate. On the contrary, helix IV adopts a non-native conformation. In the CD this is reflected in the fact that (i) the (negative) signal around 270-280 nm due to the Tyr and Phe residues in the helices I and II is conserved from the native to the intermediate while (ii) the neg-

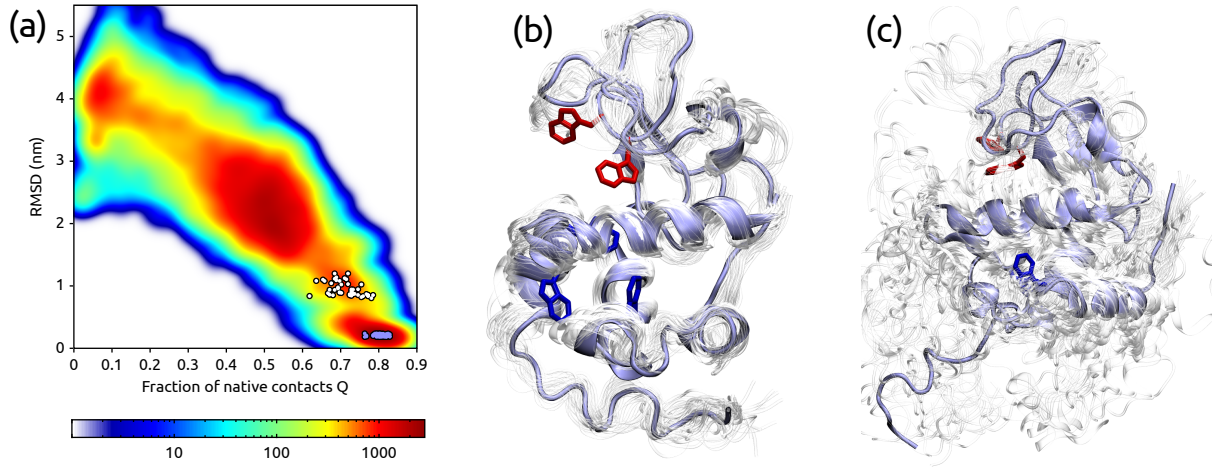

Figure 5: a) Density plot, computed using the full ensemble of rMD trial trajectories without disulfide bridges, projected onto the plane defined by the RMSD to **Native** and the fraction of native contacts Q. The high density regions correspond to long-lived intermediates. b), c) Depiction of the harvested configurations in the **Native** (b) and **I-Second** (c) basins in the case where sulfides are reduced.

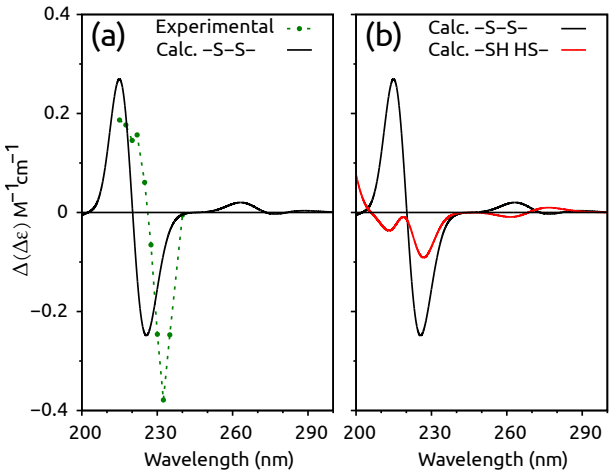

Figure 6: (a) Kinetic difference CD spectrum, calculated as **Native** minus **I-Second**, compared with the experimental kinetic difference from Ref. 47. (b) Comparison between the kinetic difference computed on the structure with and without disulfide bridges.

ative peak at ∼295 nm originated by the Trp in the helix IV is largely reduced in the intermediate.

4 Conclusion

CD has long been used to monitor the formation of secondary structures in protein folding reactions. However, CD data alone do not enable to structurally characterize the folding intermediate states with atomistic resolution, but only provide a qualitative indication of the degree of packing of the chain.

By combining enhanced MD techniques and quantum chemical simulations of excitonic spectra, here we have demonstrated that time-resolved CD signals from aromatic residues provide a unique probe of protein tertiary structure. In fact, the excitonic nature of the signals allows a direct link of the spectral evolution with even small changes in the 3D structure, thus allowing an atomistic level interpretation of the folding process.

The agreement between the theoretical predictions and the results from time-resolved CD spectra of milk canine lysozyme clearly shows the accuracy and completeness of the proposed approach. Moreover, the further application to an artificial analog where S-S bridges have been broken, evidences the high sensitivity of the method: the same number of folding intermediates is predicted but this time inconsistent results with experimental CD data are obtained. Finally, a similar study performed on a completely different protein (colicin immunity binding Im7) provided a further independent validation of the proposed analysis.

The main limitations of the approach are as-

sociated with the approximations introduced to make the calculations computationally feasible with the available computational resources. In particular, the rMD technique used to generate the trial folding trajectories in the BF approach requires in input the three-dimensional structure of the native state and is based on an arbitrary *a priori* choice of biasing coordinate. However, in a recent paper, it was shown that the model dependence associated to the choice of biasing coordinate can be eliminated by an adaptive iterative procedure, through which the reaction coordinate is computed self-consistently, at the price of increasing the computational cost by about one order of magnitude.[78]

Due to these characteristics, future applications will not be limited to the analysis of the folding of known proteins but they will deal with protein modifications and protein design with the goal of identifying unknown intermediates and predict the mechanism. As a result, a well known experimental technique such as CD will turn into a new powerful approach to elucidate the folding mechanism with fully atomistic detail.

Supporting Information Available: Details on the Bias Functional approach and the quantum chemical excitonic model. Excitonic parameters and convergence of the CD spectra. Statistical analysis of the folding trajectories without disulfide bridges. Validation study on the Im7 protein. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Acknowledgement The authors are grateful to E. Schneider who stimulated and supported this investigation. BF calculations were performed on the Tier-0 Marconi facility at CINECA.

References

(1) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *Procs. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309–1314.

(2) Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042–1046.

(3) Abaskharon, R. M.; Gai, F. *Biophys. J.* **2016**, *110*, 1924–1932.

(4) Englander, S. W.; Mayne, L. *Procs. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 8253–8258.

(5) Bhowmick, A.; Brookes, D. H.; Yost, S. R.; Dyson, H. J.; Forman-Kay, J. D.; Gunter, D.; Head-Gordon, M.; Hura, G. L.; Pande, V. S.; Wemmer, D. E.; Wright, P. E.; Head-Gordon, T. *J. Am. Chem. Soc.* **2016**, *138*, 9730–9742.

(6) Lindorff-Larsen, K.; Piana-Agostinetti, S.; Dror, R.; Shaw, D. *Science* **2011**, *334*, 517–520.

(7) Bowman, G.; Pande, V.; Noé, F. *Adv. Exp. Med. Biol.* **2013**, *797*.

(8) Onuchic, J. N.; Wolynes, P. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(9) Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K. *Biophys. J.* **1998**, *75*, 662–671.

(10) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(11) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589–1615.

(12) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58–65.

(13) Abrams, C.; Bussi, G. *Entropy* **2013**, *16*, 163–199.

(14) Vashisth, H.; Skiniotis, G.; Brooks III, C. L. *Chem. Rev.* **2014**, *114*, 3353–3365.

(15) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. *Biochim. Biophys. Acta* **2015**, *1850*, 872–877.

(16) Faccioli, P.; Beccara, S. a. *Biophys. Chem.* **2016**, *208*, 62–67.

(17) Valsson, O.; Tiwary, P.; Parrinello, M. *Annu. Rev. Phys. Chem.* **2016**, *67*, 159–184.

(18) Maximova, T.; Moffatt, R.; Ma, B.; Nussinov, R.; Shehu, A. *PLOS Comput. Biol.* **2016**, *12*, e1004619–70.

(19) Bai, Y. *Chem. Rev.* **2006**, *106*, 1757–1768.

(20) Gast, K.; Nöppert, A.; Müller-Frohne, M.; Zirwer, D.; Damaschun, G. *Eur. Biophys. J.* **1997**, *25*, 211–219.

(21) Schuler, B.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26.

(22) Sasmal, D. K.; Pulido, L. E.; Kasal, S.; Huang, J. *Nanoscale* **2016**, *8*, 19928–19944.

(23) Neuman, K. C.; Nagy, A. *Nat. Methods* **2008**, *5*, 491–505.

(24) Churnside, A. B.; Perkins, T. T. *FEBS Lett.* **2014**, *588*, 3621–3630.

(25) Michalet, X.; Weiss, S.; Jäger, M. *Chem. Rev.* **2006**, *106*, 1785–1813.

(26) Banerjee, P. R.; Deniz, A. A. *Chem. Soc. Rev.* **2014**, *43*, 1172–1188.

(27) Berova, N.; Polavarapu, P. L.; Nakanishi, K.; Woody, R. W. *Comprehensive Chiroptical Spectroscopy*; Wiley: Hoboken, NJ, 2012.

(28) Woody, R. W. *Biomed. Spectrosc. Imaging* **2015**, *4*, 5–34.

(29) Greenfield, N. J. *Nat. Protoc.* **2007**, *1*, 2876–2890.

(30) Micsonai, A.; Wien, F.; Kernya, L.; Lee, Y.-H.; Goto, Y.; Réfrégiers, M.; Kardos, J. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E3095–103.

(31) Lobley, A.; Whitmore, L.; Wallace, B. *Bioinformatics* **2002**, *18*, 211–212.

(32) Sreerama, N.; Woody, R. W. *Protein Sci.* **2004**, *13*, 100–12.

(33) Bulheller, B. M.; Rodger, A.; Hirst, J. D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2020–2035.

(34) Seibert, J.; Bannwarth, C.; Grimme, S. *J. Am. Chem. Soc.* **2017**, *139*, 11682–11685.

(35) Roder, H.; Maki, K.; Cheng, H. *Chem. Rev.* **2006**, *106*, 1836–1861.

(36) Greenfield, N. J. *Nat. Protoc.* **2009**, *1*, 2891–2899.

(37) Ranjbar, B.; Gill, P. *Chem. Biol. Drug Des.* **2009**, *74*, 101–120.

(38) Kliger, D. S.; Chen, E.; Goldbeck, R. A. In *Comprehensive Chiroptical Spectroscopy*; Berova, N., Polavarapu, P. L., Nakanishi, K., Woody, R. W., Eds.; Wiley: Hoboken, NJ, 2012; Vol. 1; Chapter 7.

(39) Hache, F. *Proc.SPIE* **2015**, *9360*, 6.

(40) Auer, H. E. *J. Am. Chem. Soc.* **1973**, *95*, 3003–3011.

(41) Woody, R. W. *Eur. Biophys. J.* **1994**, *23*, 253–262.

(42) Beccara, S.; Fant, L.; Faccioli, P. *Phys. Rev. Lett.* **2015**, *114*, 098103.

(43) Skrbic, T.; a Beccara, S.; Covino, R.; Faccioli, P. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *109*, 2330–2335.

(44) Jurinovich, S.; Pescitelli, G.; Di Bari, L.; Mennucci, B. *Phys. Chem. Chem. Phys.* **2014**, *16*, 16407–18.

(45) Padula, D.; Jurinovich, S.; Di Bari, L.; Mennucci, B. *Chem. Eur. J.* **2016**, *22*, 17011–17019.

(46) Segatta, F.; Cupellini, L.; Jurinovich, S.; Mukamel, S.; Dapor, M.; Taioli, S.; Garavelli, M.; Mennucci, B. *J. Am. Chem. Soc.* **2017**, *139*, 7558–7567.

(47) Nakao, M.; Maki, K.; Arai, M.; Koshiba, T.; Nitta, K.; Kuwajima, K. *Biochemistry* **2005**, *44*, 6685–6692.

(48) Nakatani, H.; Maki, K.; Saeki, K.; Aizawa, T.; Demura, M.; Kawano, K.; Tomoda, S.; Kuwajima, K. *Biochemistry* **2007**, *46*, 5238–5251.

(49) Paci, E.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 441–459.

(50) Camilloni, G.; Broglia, R. A.; Tiana, G. *J. Chem. Phys.* **2011**, *134*, 045105.

(51) Wang, F.; Cazzolli, G.; Wintrode, P.; Faccioli, P. *J. Phys. Chem. B* **2016**, *120*, 9297–9307.

(52) Cazzolli, G.; Wang, F.; a Beccara, S.; Gershenson, A.; Faccioli, P.; Wintrode, P. L. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 15414–15419.

(53) K. Lindorff-Larsen and S. Piana and K. Palmo and P. Maragakis and J. L. Klepeis et al., *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

(54) B. Hess and C. Kutzner and D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(55) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383–394.

(56) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835–848.

(57) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(58) Hsu, C.-P.; Fleming, G. R.; Head-Gordon, M.; Head-Gordon, T. *J. Chem. Phys.* **2001**, *114*, 3065–8.

(59) Iozzi, M. F.; Mennucci, B.; Tomasi, J.; Cammi, R. *J. Chem. Phys.* **2004**, *120*, 7029–7040.

(60) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(61) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.

(62) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–6.

(63) Sreerama, N.; Woody, R. W. *Methods Enzymol.* **2004**, *383*, 318–351.

(64) Arulmozhiraja, S.; Coote, M. L. *J. Chem. Theory Comput.* **2012**, *8*, 575–584.

(65) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

(66) Jurinovich, S.; Cupellini, L.; Guido, C. A.; Mennucci, B. *J. Comp. Chem.* **2018**, *39*, 279–286.

(67) Koshiba, T.; Yao, M.; Kobashigawa, Y.; Demura, M.; Nakagawa, A.; Tanaka, I.; Kuwajima, K.; Nitta, K. *Biochemistry* **2000**, *39*, 3248–3257.

(68) Loco, D.; Jurinovich, S.; Di Bari, L.; Mennucci, B. *Phys. Chem. Chem. Phys.* **2016**, *18*, 866–877.

(69) Jurinovich, S.; Viani, L.; Prandi, I. G.; Renger, T.; Mennucci, B. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14405–14416.

(70) Cupellini, L.; Jurinovich, S.; Campetella, M.; Caprasecca, S.; Guido, C. A.; Kelly, S. M.; Gardiner, A. T.; Cogdell, R.; Mennucci, B. *J. Phys. Chem. B* **2016**, *120*, 11348–11359.

(71) Mizuguchi, M.; Arai, M.; Ke, Y.; Nitta, K.; Kuwajima, K. *J. Mol. Biol.* **1998**, *283*, 265–277.

(72) Sasahara, K.; Demura, M.; Nitta, K. *Biochemistry* **2000**, *39*, 6475–6482.

(73) Glättli, A.; Daura, X.; Seebach, D.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2002**, *124*, 12972–12978.

(74) Brkljača, Z.; Čondić-Jurkić, K.; Smith, A.-S.; Smith, D. M. *J. Chem. Theory Comput.* **2012**, *8*, 1694–1705.

(75) Woody, R. W.; Dunker, K. A. In *Circular Dichroism and the Conformational Analysis of Biomolecules*; Yang, J. T., Fasman, G. D., Eds.; Plenum Press: New York, NY, 1996; Chapter 4, pp 109–157.

(76) Spence, G. R.; Capaldi, A. P.; Radford, S. E. *J. Mol. Biol.* **2004**, *341*, 215–226.

(77) Li, Z.; Hirst, J. D. *Chem. Sci.* **2017**, *8*, 4318–4333.

(78) Orioli, S.; a Beccara, S.; Faccioli, P. *J. Chem. Phys.* **2017**, *147*, 064108.

Graphical TOC Entry