Energetic self-folding mechanism in α-helixes

Adolfo Bastida,* José Zúñiga, Alberto Requena, and Javier Cerezo*

Departamento de Química Física. Universidad de Murcia, 30100 Murcia, Spain

E-mail: bastida@um.es; jcb1@um.es

Abstract

A novel energetic route driving the folding of a polyalanine peptide from an extended conformation to its α-helix native conformation is described, supported by a new method to compute mean potential energy surfaces accurately in terms of the dihedral angles of the peptide chain from extensive Molecular Dynamics simulations. The Energetic Self-Folding (ESF) route arises specifically from the balance between the intrinsic propensity of alanine residues towards the α_R conformation and two, opposite, effects: the destabilizing interaction with neighbor residues and the stabilizing formation of native hydrogen bonds, with the latter being dominant for large peptide lengths. The ESF mechanism provides simple but robust support to the nucleation-elongation, or zipper models, and offers a quantitative energetic funnel picture of the folding process. The mechanism is validated by the reasonable agreement between the computed folding energies and the experimental values.

1 Introduction

The study of protein folding has revealed some remarkable paradoxes, the best-known probably being due to Levinthal,[1,2] who argued that the conformational space of a protein is so huge that the protein would require an almost unlimited amount of time to find its native structure by random search. The solution of the paradox seems to be obvious, pointing to the fact that there must be specific folding pathways which drive the protein towards its native structure. These folding pathways are conceived as trajectories in the free energy landscape and account for the protein conformations as a function of the degrees of freedom of the system, such as the dihedral angles along the peptide backbone.[1,2] During the folding process, the conformational entropy of the protein decreases since the formation of native contacts reduces the accessible conformational space. This entropic reduction must be compensated by the remaining contributions to the free energy, which are the energy resulting from the intra- and intermolecular interactions and the entropy of the solvent, for the total free energy to decrease during the folding process. This thermodynamic picture is known as the folding funnel.[1–5]

At the atomistic level, the development of reliable Molecular Mechanics force fields has enabled successful simulations of the folding of small proteins by Molecular Dynamics (MD).[6,7] Although the extension of the atomic simulations to more complex proteins is limited by the increasing computational effort required, the availability of more accurate force fields and optimized MD packages has resulted in atomistic simulations being very useful tools in this respect.[2,8–10] It is, however, paradoxical that the access to precise atomic simulations has not substantially improved our knowledge of the general principles governing the routes and speed of the protein folding.[6] These principles should fill the gap between the atomistic and thermodynamic pictures,[7] thus allowing the prediction of the protein's native structure from its aminoacid sequence.[6,11]

The problem of extracting the folding patterns of proteins from MD simulations arises from the intrinsic complexity of the system, in which the different energetic and entropic contributions find a delicate balance that minimizes the total free energy along the folding process. Indeed, as folding progresses, both the potential energy and the entropy of the protein decreases, although resulting in opposite effects on the free energy, whereas the entropy of the solvent plays a much more uncertain role.[6,12] The global structure of the protein is determined by the conformation of each residue, which is defined in turn by the backbone dihedral angles ϕ and ψ (see Figure 1**a**). A deep understanding of the folding process therefore requires accurate estimations of the energetic and entropic contributions to the free energy in terms of the dihedral coordinates.

Calculation of the free energy[13,14] is usually performed using the mean force potential along some coordinates obtained by directly counting the times that the molecule visits a given conformational subregion, which sometimes requires enhanced sampling techniques.[15,16] The resulting free energy landscapes are intrinsically rugged,[11,17] and typically include different local minima, β, pPII, α_{R},. . . , which act as traps during folding,[18,19] with the free energy barriers being of the same order of magnitude as the thermal energy, so the molecule may overpass them during the folding process. Accordingly, the free energy landscape must be calculated with an accuracy to within $k_{\mathrm{B}}T$ (0.59 kcal/mol) for it to be considered physically meaningful.

The calculation of the energetic and entropic contributions to the free energy separately from atomistic simulations is much more problematic,[14] with existing methods having an often unsatisfactory efficiency/accuracy ratio. In principle, the energetic contribution should be easier to evaluate since it is given by just the potential energy differences between the distinct conformations. Computing the potential energy of the i-residue in terms of its dihedral angles ϕ_i and ψ_i implies averaging the potential energy over the conformations of all the remaining residues. In practice, such a calculation is not generally possible, due to the numerical uncertainties arising from the huge number of bonded and, more significantly, non-bonded interactions to be evaluated.

In this work, we present a novel effective method, called the K2V method, to calculate the mean (ϕ_i, ψ_i) potential energy landscapes from MD simulations, to gain insights on the energetic contributions of the free energy to the folding process. These landscapes allow us to go beyond the two state (folded and unfolded) model, in which the energetic changes due to the folding are characterized by a single quantity, ΔH or ΔE, depending on the choice of the NPT or NVE ensembles,[7,20] although both quantities are basically identical in water solutions at room temperature, a model which is indeed at odds with the image of a rugged free energy landscape. The evaluation of the potential energy landscapes at different steps during the folding process has allowed us, in turn, to unravel the presence, or absence, of energy folding paths guiding the search for the native structure, and to investigate the role of the intramolecular hydrogen bonds (HBs) in the energetic stabilization of the α-helix.

Polyalanine peptides show a remarkable α_{R}-helix propensity,[19] which justifies its use in theoretical investigations of conformational processes. As model systems, we have chosen, accordingly, the capped polymers of alanine (ACE-(Ala)$_m$-NME) with m=1, 2, 3, 4, 6, 8 and 10. The $m = 10$ molecule, usually referred to as deca-alanine, has often served as a methodological proof of concept,[21] since it provides a good model of larger peptides, and so it is used in this study. Deca-alanine is also long enough to offer a good representation of the formation of the α_{R}-helix, and short enough to allow the extensive MD simulations necessary to reach convergence in the statistical analysis, and to reduce the presence of misfolded intermediates that may significantly alter the folding rates.[19] We should note that our attention is essentially focused on the search for energetic mechanisms that may assist the propagation of the elements of the secondary structure. Therefore, we concentrate our analysis on the initial folding process, conducting the simulations up to 15 ns, a long enough period of time to run from a mostly extended conformation to a structure where the percentage of residues in the α_{R} conformation is significant, although the conformational equilibrium had not been yet reached.

2 Methods

2.1 MD Simulations.

MD simulations of the ACE-(Ala)$_m$-NME molecules with m =1, 2, 3, 4, 6, 8 and 10, dissolved in water, were carried out using the GROMACS package v2016.4.[22,23] Each solute molecule was surrounded by a number of water molecules ranging from 600 to 2000 (depending on the length of the polyalanine chain) and placed in a cubic box of a size chosen to reproduce the experimental density of the liquid at room temperature. All the polyalanine molecules were described using the CHARMM27[24,25] force field. Two additional sets of independent simulations were carried out for deca-alanine using the AMBERGS[26] and the OPLS-AA[27] force fields. The water solvent was in all cases described using the flexible TIP3P model. Periodic boundary conditions were imposed in the simulations using the Particle-Mesh Ewald method to treat the long range electrostatic interactions. The equations of motion were integrated using a time step of 0.5 fs.

Since we were mainly interested in the first steps of the folding, the chronology of the MD simulations for every polyalanine molecule and force field was as follows. The system was first equilibrated during 0.4 ns in a NVT ensemble at 298 K by coupling to a thermal bath. In this period of time, the deca-alanine molecule was kept frozen in a fully extended conformation, with all dihedral angles fixed at the β conformer maximum. Then, the velocities of the polyalanine atoms were reset randomly using a Boltzmann distribution, after which an additional equilibration of 40 ps followed. The last process was repeated 112 times with different sets of velocities. Each of these 112 initial configurations were propagated for 15 ns, generating the same number of distinct trajectories. During these production runs, the kinetic energy of the deca-alanine atoms, the total potential energy of the system, and the values of the dihedral angles of the ten residues were written every 5 fs. This computational strategy allowed us to obtain thermally equilibrated systems, with the dihedral angles of the molecule populating different regions of the conformational space. Throughout the MD simulations, the number of hydrogen bonds (HBs) was computed using the criteria already adopted by Zewail and co-workers.[19]

2.2 The K2V algorithm.

The calculation of the mean potential energy maps of a given residue in terms of its dihedral angles by averaging the total potential energy of the system calculated for the remaining dihedral conformations is not viable practically for the following reasons: (1) the huge statistics necessary to cancel the mean contributions of the remaining degrees of freedom, including the solvent ones, (2) the accuracy with which the potential energy is evaluated, probably not high enough to reproduce the comparatively slight potential energy differences corresponding to small variations of the dihedral angles, and (3) the numerical errors raising in the evaluation of the average potential energy values due to the vast quantity of data involved in the calculations.

We decided, then, to explore the possibility of determining the potential energy maps using the kinetic energy of the atoms. This idea has its origins in the fact that the changes in the kinetic energy of any internal coordinate of the system in a short time interval δt occur in two ways, by variation of the potential energy function due to the internal coordinate displacements, and from the energy fluxes with the remaining degrees of freedom driven by the kinetic and/or potential couplings. If the system is at thermal equilibrium, the average contribution of the energy fluxes vanishes since the average kinetic energy per degree of freedom must be equal to $k_\mathrm{B}T/2$. Accordingly, the average changes of the kinetic energy of the system in going form conformation i to other conformation j $(i \to j)$ is expected to be equal to minus the potential energy difference, i.e.,

$$\overline{\Delta T}_{i\to j} = -\overline{\Delta V}_{i\to j} \tag{1}$$

where the averaged kinetic (ΔT) and potential (ΔV) energy differences run over all the transitions taking place along the trajectories.

By using the kinetic energy changes between conformations, we get the crucial advantage of accurately expressing the kinetic energy as the sum of individual atomic contributions, something that cannot be done for the potential energy. This allows us to select the atoms which have a real kinetic energy contribution to the displacement of a given internal coordinate. Thus we find that, for the deca-alanine molecule, only a few atoms really have an influence on the motion of the dihedral angles, and that these atoms always occupy the same relative positions with respect to the dihedral angles (see Supplementary Figure 1). As expected, the atoms with the highest contributions are those pertaining to the dihedral angles, that is, the C_{i-1}-N_i-$C_{\alpha,i}$-C_i and N_i-$C_{\alpha,i}$-C_i-N_{i+1} backbone atoms for ϕ_i and ψ_i, respectively. These atoms together account for ~97% of the total kinetic energy involved in the displacements, which increases up to ~99% when the contribution of the $H_{\alpha,i}$ atoms is added (See Supplementary Table 1). We therefore use the kinetic energy of these six atoms, referred to as the CNCHCN atoms, to unravel the mean potential energy as a function of the dihedral angles of each residue.

In order to describe the algorithm employed to calculate the potential energy maps, the so-called K2V algorithm, let us consider a 2D (ϕ, ψ) map divided in a squared $N \times N$ grid. The value of N is chosen to be high enough for the potential energy to be considered constant in every square of the grid. These constant values are labeled as V_1, V_2,$\ldots$,V_{N^2} and are the unknowns to be determined. Using the coordinates and the kinetic energies of the atoms exported at δt time intervals, we evaluate the mean change of the kinetic energy of the CNCHCN atoms which are in the ith point of the grid at time t when passing to the jth point at time $t + \delta t$ as follows:

$$\overline{\Delta T}_{i \to j} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} \left(T_j^{(k)}(t + \delta t) - T_i^{(k)}(t) \right) \tag{2}$$

where n_{ij} is the number of $i \to j$ conformational changes taking place in the simulation. A regression analysis is then performed to calculate the V_i values by minimizing the weighted sum of squared residuals given by

$$R^2 = \sum_{i,j}^{N^2} n_{ij} \left(\overline{\Delta T}_{i \to j} + (V_j - V_i) \right)^2 \tag{3}$$

The condition

$$\frac{\partial R^2}{\partial V_i} = 0, \quad i = 1, \ldots, N^2 \tag{4}$$

leads accordingly to the following system of linear equations

$$V_i \sum_{\substack{j \neq i \\ j=1}}^{N^2} (n_{ij} + n_{ji}) - \sum_{\substack{j \neq i \\ j=2}}^{N^2} (n_{ij} + n_{ji}) V_j = \sum_{j=1}^{N^2} (n_{ij} \overline{\Delta T}_{i \to j} - n_{ji} \overline{\Delta T}_{j \to i}), \quad i = 2, \ldots, N^2 \tag{5}$$

whose solution provides the V_i values. We note that the value $V_1 = 0$ is arbitrarily chosen to set energy origin.

In practice, not all the dihedrals conformations are explored by the molecule, due to energetic constrains, so the real number of V_i variables is much smaller than N^2. Typically, we find that $\sim$40% of the conformational space is not visited, which reduces the number of V_i variables to be calculated appreciably. The choice of the grid resolution (N) deserves some additional comments. In principle, high values of N are desirable in order to increase the resolution of the potential energy maps. However, this has two practical limitations. The first is about the number of data available from the simulations, since higher resolutions require more data to evaluate accurately the mean values of the kinetic energy differences between the grid points. The second limitation has to do with the dimension of the linear equations system to be solved. Grids with N in the range 90-180 imply solving equation systems with 5000 to 20000 variables. Higher dimensions would require special computers with extended memories. The maps shown in the present work have a 90×90 resolution and are calculated using a computer with 32 Gb of RAM. This means that every grid square has a $4^\circ \times 4^\circ$ dimension. This size imposes a lower limit on the choice of the elapsed time (δt) used to evaluate the kinetic energy differences. The algorithm is effective only when the displacement of the dihedral angles during the δt time interval is of the same magnitude as the size of the grid cell, ordinarily involving, therefore, a transition between different grid cells. Simultaneously, the value of δt cannot be too long since, otherwise, the validity of the relation between the changes of kinetic and potential energies implied in equation (1) would be lost due to the change of the atomic environment. The value chosen in the present work, $\delta t = 5$ fs, is the results from a compromise. The average displacements of the dihedral angles are $\sim 3^\circ$, which assures that most of the displacements are going, to modify the grid point while they are short enough to maintain the correlation in the changes of the kinetic and potential energies. In the Supplementary Information we describe the tests carried out to evaluate the performance and reliability of the K2V algorithm (see Supplementary Figures 2, 3 and 4).

3 Results and Discussion

We start our discussion by considering the results obtained from the MD simulations for the deca-alanine molecule. In the Ramachandran plot shown in Figure 1**a** (see also Supplementary Figures 5 and 6) we observe the presence of five local maxima, corresponding to the β, pPII, α_{R}, α', and α_{L} conformers (see Supplementary Table 2 for a precise location of the conformational maxima). In order to quantify the time evolution of the conformational changes, we have collected the conformational populations within $\pm 10^\circ$ intervals around each maximum. We should note that the Ramachandran plot shown in Figure 1**a**, as well as the remaining plots presented in this work, are averaged over all the residues of the molecules, since the differences between residues are in general of little significance for the present study. Also, it should be emphasized that the Ramachandran plot shown in Figure 1**a** does not correspond to the conformational equilibrium, since the conformation of the deca-alanine molecule evolves during the 15 ns simulations, as observed in Figure 1**b**. Initially, only the β and pPII conformations have a significant contribution of around $\sim$10%, while $\sim$75% of the dihedral angles are in none specific conformational regions. During the propagation, the population of the α_R conformers increases monotonically and reach a value close to 40%.

In Figure 2**a** (see also Supplementary Figures 7 and 8) we show the mean potential energy map of deca-alanine as a function of the dihedral angles, obtained using the K2V algorithm (see section 2.2). This map corresponds to the early stages of the simulations, when none of the remaining residues is in the α_R conformation, and shows five minima that are located in the same positions as the five maxima in the Ramachandran plot in Figure 1**a**. The deepest minimum corresponds to the α_R conformer, and is followed by the pPII, α', α_L and β minima in order of increasing energy. The potential energy wells are separated from each other by potential energy barriers of different magnitudes, with the lowest barriers being those between the α_R-α' and α'-pPII conformations. In general, the potential surface looks rough, as expected. Typically, the barriers between the conformers are a few orders of magnitude higher than the thermal energy ($k_\mathrm{B}T$), high enough to trap the residue in a given conformation for an elapsed time, although they can be surpassed when thermal fluctuations concentrate enough energy in the dihedral degrees of freedom.

Inspection of the trajectories reveals that the mean energy map of Figure 2**a** is not static but evolves during the folding process, indicating that the forces acting on the dihedral angles of a given residue depend on the conformation of the other residues. This effect is better shown by computing the potential energy maps filtering the data in order to ensure that a certain amount of global α_R conformation is within the chain. In particular, we have calculated separated potential energy maps for a given residue of the polypeptide as a function of the number, N_{α_R}, of the remaining residues of the molecule being in the α_R conformation. In Figure 2**b** we present these maps for N_{α_R} varying from 0 to 6, and they show notable changes in the relative depths of the potential energy wells with N_{α_R} which gradually favor the α_R conformation. Interestingly, the potential energy minima of the pPII, α', α_L and β conformers relative to the α_R minimum increase all of them linearly with N_{α_R} and have similar slopes (see Supplementary Figures 9, 10 and 11). On the other hand,

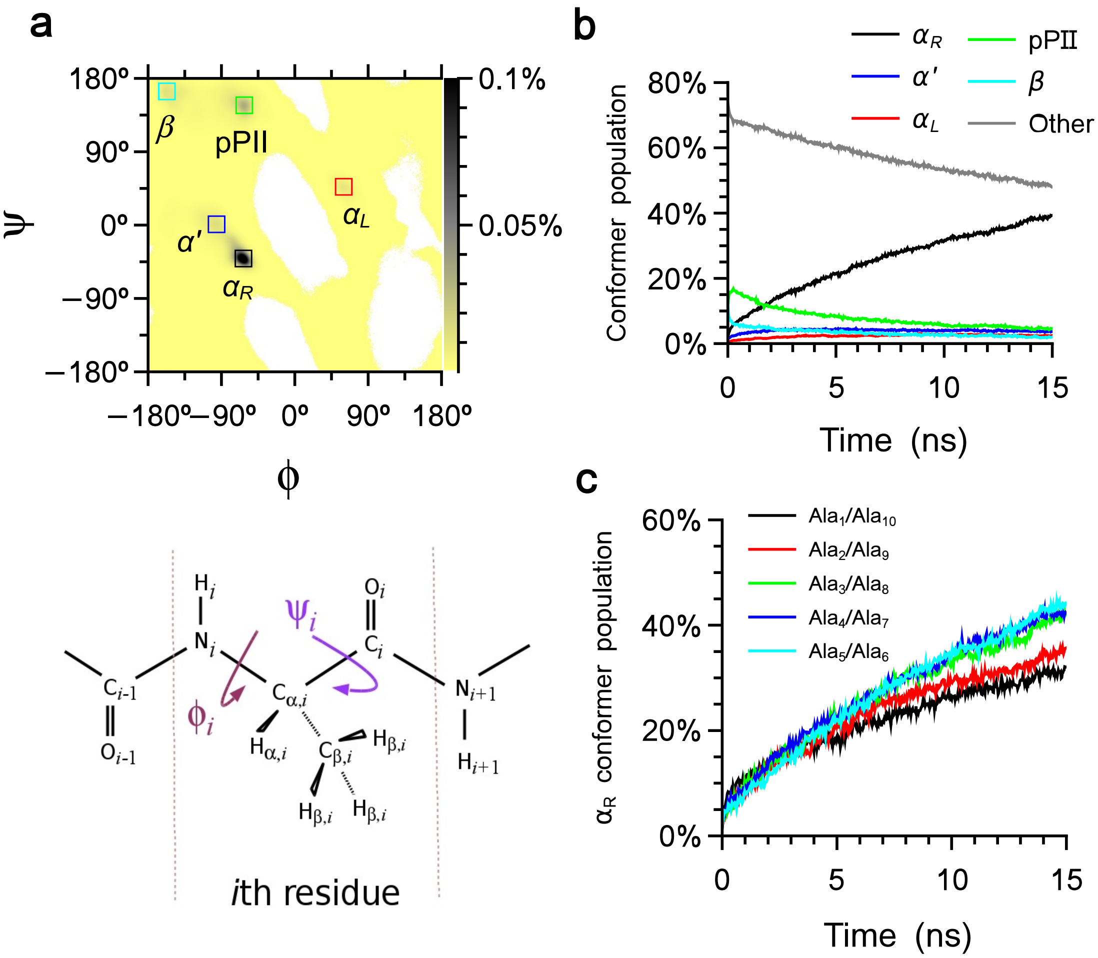

Figure 1: Results for the deca-alanine molecule. **a.** Ramachandran plot. The five conformational regions are defined with ±10° intervals around each maxima. **b.** Time evolution of the conformer populations. **c.** Time evolution of the α_R conformer populations for the residues of deca-alanine paired according to their place with respect to the center of the molecular chain. Results obtained using the CHARMM27 force field.

outside the α_R region, the potential energy map is practically unaffected by N_{α_R}, as better illustrated by the minimum energy paths between the β and $\alpha_{\rm R}$ conformations included in Figure 2**c** (see also Supplementary Figure 12). Overall, Figure 2**b** reveals the existence of an energetic funnel that favors the successive formation of $\alpha_{\rm R}$ residues as the folding process advances.

The stabilization of the $\alpha_{\rm R}$ conformers is expected to have a direct impact on the folding energies of the residues. Experimentally, the folding energy is obtained by assuming a two-state ($\alpha_{\rm R}$ folded and unfolded) model,[28,29] so we have evaluated it using the same prescription, that is, by dividing the dihedral conformational space into a folded and an unfolded region. The folded region is identified with the $\alpha_{\rm R}$ region centered at (-63°, -41°) with ±10° intervals, whereas the unfolded region contains all other molecular conformations. Using the potential energy maps for a given value of N_{α_R} and the Ramachandran plots, we simply calculate the mean values of the potential energy inside and outside the folded region, and the difference gives the folding energy as a function of N_{α_R}. In order to disentangle the effects that modulate the folding energy, we have also computed it for the series of the polyalanine molecules of increasing size (m = 1 to 10). Only the results for $N_{\alpha_{\rm R}}$ conformations which account for at least 5% of the total are used in the analysis to assure statistical reliability.

Figure 3 shows the computed folding energies for the different polyalanine chains as a function of $N_{\alpha_{\rm R}}$. As observed, the folding energy varies with $N_{\alpha_{\rm R}}$ following a trend that strongly depends on the size of the polyalanine chain. In order to rationalize these results, let us consider first the folding energy for the $m = 1$ molecule, which is -0.83 kcal/mol. Since this molecule only contains one residue, the value of the folding energy indicates that there is an intrinsic stabilization of the folded conformation with respect to the unfolded conformers, even in the absence of interactions with other residues. This stabilization comes, therefore, from the molecular structure of the alanine residue. For the two residues molecule, $m = 2$, the folding energy per residue reduces to $\sim$0.4 kcal/mol when $N_{\alpha_{\rm R}}$ increases from 0 to 1, that is, when the other residue folds into the $\alpha_{\rm R}$ conformation. So there is a neighborhood

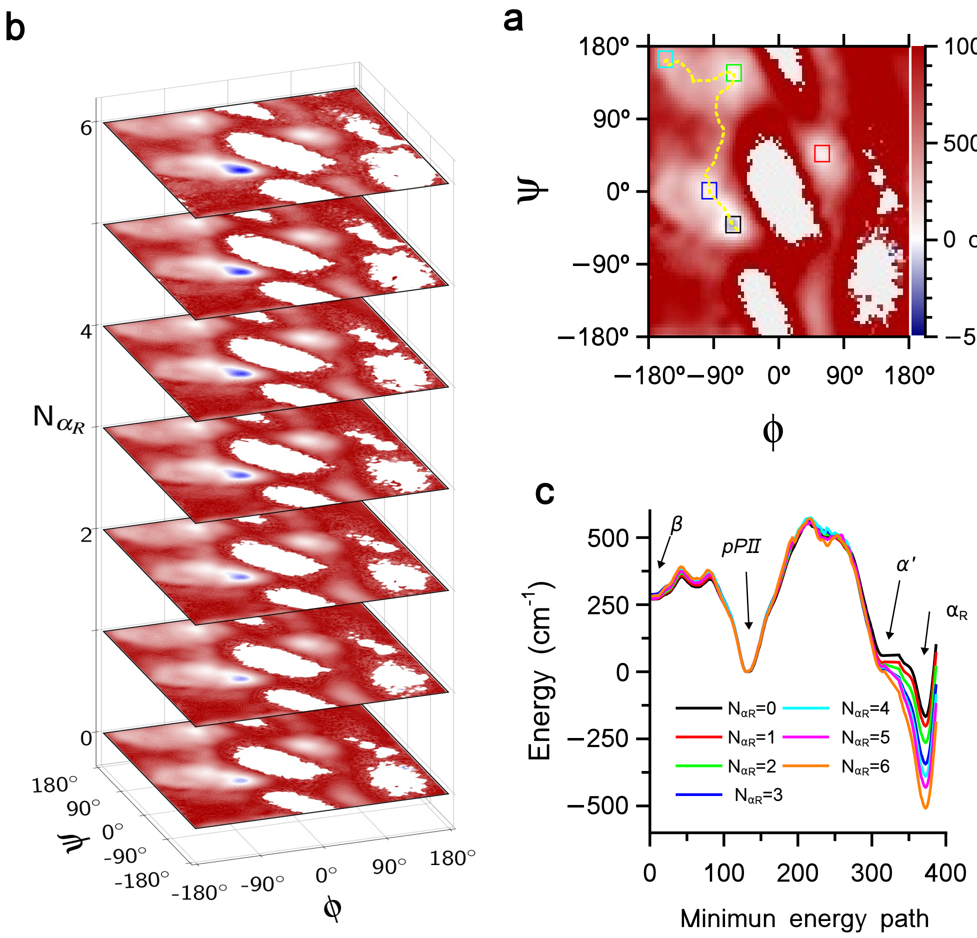

Figure 2: Results for the deca-alanine molecule. **a.** Mean potential energy map in terms of the dihedral angles (ϕ, ψ) computed for mostly extended conformations when none of remaining residues was in the $\alpha_{\rm R}$ conformation. The five minima (also observed along all the dynamics) are depicted with colored squares, and the yellow dashed line traces to the minimun energy path. **b.** Mean potential energy maps as a function of the number of alanine residues of the deca-alanine molecule in the $\alpha_{\rm R}$ conformation ($N_{\alpha_{\rm R}}$). **c.** Potential energy along minimum energy path between the β, pPII, α', and $\alpha_{\rm R}$ conformers for the $N_{\alpha_{\rm R}}$=0-6 maps. Results obtained using the CHARMM27 force field. The origin of energy is arbitrarily placed at the minima of the pPII conformer.

effect that hinders the folding of the adjacent residues when a given residue is already in the α_{R} conformation. It is worth noting that the effect of neighbor residues on the conformational flexibility of a given residue has already been observed by analyzing the dynamics of polypeptide chains.[30–32] The neighboring residue effect also diverts the energetic balance in peptide studies from group additivity, that is, from the isolated pair hypothesis.[33] The results obtained for the $m = 3$ polyalanine reveal that the folding energy for the $N_{\alpha_{\mathrm{R}}} = 0$ conformers is more negative than the corresponding energy for the shorter molecules. Also, for $m = 3$ alanine the increase of the folding energy when $N_{\alpha_{\mathrm{R}}}$ goes from 0 to 1 is much smaller than that for $m = 2$ alanine. This change of tendency is confirmed in the $m = 4$ molecule, for which the folding energy decreases as $N_{\alpha_{\mathrm{R}}}$ increases, and results from the formation of an increasing number of intramolecular HBs that stabilize the α_{R} conformations of the residues (see Supplementary Table 3). In fact, the increasing number of native $(i, i+4)$ HBs, which can formed only in the molecules with $m \geq 4$, drives the folding energies towards even more negative values.

The neighborhood effect and the formation of HBs have opposite effects, both intensifying as $N_{\alpha_{\mathrm{R}}}$ increases, and the balance between them may well be responsible for the variation of folding energy of the different polyalanine molecules. Interestingly, HBs dominate for longer chains (m =6, 8 and 10), giving folding energies which decrease almost linearly with the number of residues in the α_{R} region. The slopes of these linear variations scale approximately with $1/m$, since the calculated folding energies are averaged for the m residues of the molecule. Assuming that the three effects, namely the intrinsic stabilization of α_R conformations, the neighborhood effect and the HBs formation, have unconnected contributions, the folding energy as a function of the conformation, characterized by $N_{\alpha_{\mathrm{R}}}$, and the length of the molecule, m, can be modeled as follows

$$\Delta E_{\mathrm{f,residue}}(m, N_{\alpha_{\mathrm{R}}}) = E_{\mathrm{int}} + n_{\mathrm{nnr}}(m, N_{\alpha_{\mathrm{R}}})\, E_{\mathrm{nnr}} + \frac{f}{m}\, n_{\mathrm{HB}}(m, N_{\alpha_{\mathrm{R}}})\, E_{\mathrm{HB}} \qquad (6)$$

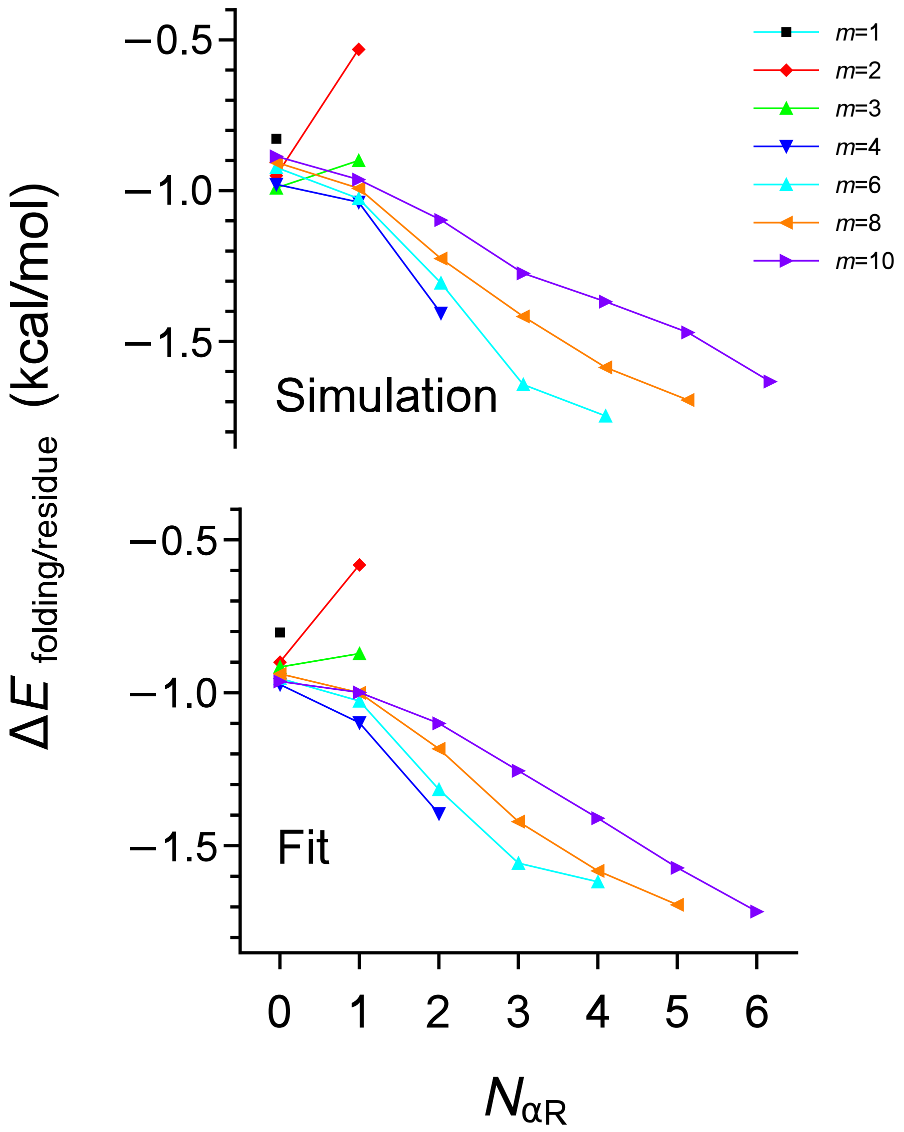

Figure 3: Folding energies as a function of the number of alanine residues of the polyalanine molecules (ACE-(Ala)$_m$-NME) in the α_R conformation (N_{α_R}). **a** Results obtained from the MD simulations. **b** Fits to Equation (6).

where E_{int}, E_{nnr}, and E_{HB} are, respectively, the energy values of the intrinsic, nearest neighbor residues, and HB contributions, and n_{nnr} and n_{HB} are the average numbers of nearest neighbor residues in the α_{R} conformation and of HBs, whose dependence on m and $N_{\alpha_{\mathrm{R}}}$ is explicitly indicated. The n_{nnr} and n_{HB} values are extracted from the simulations, as detailed in Table 3 of the Supplementary. The f/m factor accounts for the number of residues affected by each HB. For the shorter polyalanine molecules, with $m \leq 3$, f is equal to m, while for the longer chain molecules f should be close to but smaller than 4, since inner residues can participate in more than one HB. The fits of the folding energies to Equation (6) are shown in Figure 3**b** and the values of the parameters fitted are $E_{\mathrm{int}} = -0.80$ kcal/mol, $E_{\mathrm{nnr}} = 0.33$ kcal/mol, $E_{\mathrm{HB}} = -0.80$ kcal/mol and f =3.4. The values of E_{int}, E_{nnr}, and f are consistent with the estimations discussed above, and the HBs contribution is also in line with previous estimations of the HBs energy for proteins in solution.[34]

It is noteworthy that the patterns of the folding energies shown in Figure 3 can be reasonably reproduced using simple concepts, despite the apparent complexity of the simulations performed to obtain them. In this sense, our analysis reveals the existence of an energetic self-folding (ESF) mechanism resulting from the superposition of the three effects considered, which guides the polyalanine molecules towards the formation of the α_{R}-helix. We should note that the values obtained for the folding energies in the longer molecules are between 1 and 3 times higher than thermal energy at room temperature. This means that the energetic gains from the ESF mechanism plays an important role in the folding process, independently of possible additional entropic contributions, and of the mechanism not being completely deterministic. Indeed, at certain times, the ESF mechanism can be counteracted by singular thermal fluctuations of the kinetic energy of the atoms involved in the displacement of the dihedral angles. The folding path of a molecule results, therefore, from the balance between the ESF mechanism and the thermal fluctuations.

The ESF mechanism also provides a simple explanation for the fact that the folding of the inner residues is favored with respect to the folding of the outer residues, as has been shown in previous simulations[19] and experimentally.[35–37] Figure 1**c** (and Supplementary Figures 13 and 14) shows the time-evolution of the $\alpha_{\rm R}$ conformer population percentage for deca-alanine residues, extracted from the simulations. For clarity, we have grouped them according to their relative position with respect to the center of the molecular chain. As observed in this figure, the folding of the $\mathrm{Ala}_1/\mathrm{Ala}_{10}$ residues is comparatively disfavored, followed by the folding of the $\mathrm{Ala}_2/\mathrm{Ala}_9$ residues. This effect decreases significantly in magnitude as we move toward the center of the molecule, with the differences between the Ala_4/Ala7 and $\mathrm{Ala}_5/\mathrm{Ala}_6$ pairs of residues being practically negligible and absorbed by the statistical fluctuations. If we analyze these results using the ESF mechanism, we first note that the $\mathrm{Ala}_1/\mathrm{Ala}_{10}$ residues have only one nearest neighbor residue and can participate in one HB; and, second, that the remaining residues have two nearest neighbor residues but can participate in a higher number of HBs, the further away they are from the extremes: 2 HBs for the $\mathrm{Ala}_2/\mathrm{Ala}_9$ residues, 3 for the $\mathrm{Ala}_3/\mathrm{Ala}_8$ and 4 for the rest ones. Since the energetic stabilization provided by the HBs is higher in magnitude than the destabilization caused by the neighbor residues in $\alpha_{\rm R}$ conformations, the net effect is that the folding process is progressively favored from the extreme residues up to the $\mathrm{Ala}_4/\mathrm{Ala}_7$ residues, as observed.

An additional test of the reliability of the results obtained from the simulations can be performed by calculating the average folding energy per residue directly from the deca-alanine molecule, assuming that the linear variations shown in Figure 3 (see also Supplementary Figure 15) hold for the higher values of $N_{\alpha_{\rm R}}$. The averaged folding energy per residue thus obtained using the CHARMM27, AMBERGS, and OPLS-AA force fields are, respectively, -1.42, -1.23, and -1.26 kcal/mol. It is notable that such similar values are obtained from the three different force fields, being the CHARMM27 value slightly more negative due to a comparatively higher stabilization of the $\alpha_{\rm R}$ well (see Supplementary Figures 9, 10, and 11. In addition, experimental findings give folding energies between -0.9 ± 0.1[28] and -1.0 ± 0.1[29] kcal/mol per residue for the alanine peptide helix, which are ∼25% higher than the folding energies per residue calculated theoretically. We find these differences to be acceptable within the usual degree of accuracy provided by the molecular mechanics force fields and the discrepancies expected between the theoretical model and the actual experimental setup. In this respect, for instance, the interpretation of the experimental measurements is made under the assumption of the initial equilibrium percentage of α_{R} residues, a quantity that is not easy to establish,[29] while our simulations are not performed at conformational equilibrium conditions. Taking into account also that the calculated folding energies are obtained using exclusively Ramachandran plots and potential energy maps, we think that the reasonable agreement provided by our theoretical model with the experimental findings significantly supports the methodologies used in this work. We finally note that a value for the average folding energy similar to those given above for deca-alanine is obtained for the $m = 8$ polyalanine molecule in our simulations, in agreement with the experimental observation that the enthalpy of helix formation per residue, is independent of the peptide length.[28]

4 Conclusions and perspectives

In this work, we have identified an energetic route guiding the folding towards α-helix in polyalanine chains by means of extensive MD simulations. This route basically consists of the stabilization of the α_R conformation of a given residue when the adjacent residues are also in the α_R conformation, an effect which we refer to as an energetic self-folding (ESF) mechanism. The evidence that supports this mechanism comes from the analysis of the potential energy landscape associated to the (ϕ_i, ψ_i) dihedral angles describing a given residue, which determines its conformation as function of the number N_{α_R} of remaining residues that are in the α_R conformation. The results obtained for deca-alanine reveal a stabilization of the α_R conformations as N_{α_R} increases, which has, in turn, a direct impact on the folding energy per residue by causing it to decreases linearly with N_{α_R}. The variation of the folding energy with N_{α_R} depends specifically on the balance between the intrinsic stabilization of the α_R conformation within the alanine residue, which is independent of

N_{α_R}, and two opposed effects that increase with N_{α_R}, which are the interaction with neighbor residues, which destabilizes the α_R conformation, and the formation of native HBs, which stabilizes the folding. Such effects are individualized and quantified by computing the folding energy as a function of N_{α_R} for polyalanine chains of increasing size and for deca-alanine. Our model provides an intrinsic energy stabilization of $-0.80\,\text{kcal/mol}$, a destabilization energy per neighbor residue of $0.33\,\text{kcal/mol}$, and a stabilization energy per HB of $-0.80\,\text{kcal/mol}$. Interestingly, the HB stabilization dominates for long chains, leading to a nearly linear decrease of the folding energy as function of N_{α_R}. For large polyalanine peptides, the slope of these variations is inversely proportional to the chain length which, in turn, provides average folding energies independent of the peptide length.

A key feature in our work is the ability to compute the energy landscapes for the dihedral angles of the residue, ϕ_i and ψ_i, averaged over the remaining degrees of freedom, from the MD simulations. This non-trivial problem is accomplished by introducing a completely new methodological tool, the K2V method, based on the computation of the kinetic energy of a selected set of atoms to evaluate the changes in the potential energy. This methodology, applied here for the first time, provides converged potential energy maps and, when coupled to data filtering schemes in terms of N_{α_R}, delivers individual maps for different values of N_{α_R}, which give fundamental support to our findings.

In our view, the results presented in this work give stimulating clues to rationalize the folding process from an energetic point of view. This work marks in a way the beginning of a new line of research that promises to deliver key pieces in the solution of the intricate puzzle of protein folding dynamics. On the one hand, the origin of the ESF mechanism needs to be understood at a molecular level, which would require an exhaustive analysis of both, the intramolecular and intermolecular (non-bonded) energetic contributions to the free energy. Interestingly, the ESF effect seems to be closely connected to the correlation of the neighboring residues conformations, a key concept usually invoked to explain the small perturbation of the polypeptide chain caused by the conformational oscillations within a residue. On the other hand, the K2V method is readily applicable to more complex polypeptides, including arbitrary aminoacid sequences, thus opening the door to investigate the feasibility of the ESF mechanism in terms of the aminoacids nature, and its role in the folding propensity to specific motifs. Another appealing point of the K2V method is that, when accompanied by accurate computation of the free energy, which is currently viable, it may unlock the evaluation of the entropic contributions of both the solute and the solvent, which would contribute to clarifying the folding dynamics further.

References

(1) Dill, K.; Chan, H. From Levinthal to pathways to funnels. *Nature Struc. Biol.* **1997**, *4*, 10–19.

(2) Li, B.; Fooksa, M.; Heinze, S.; Meiler, J. Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Critical Review in Biochem. and Mol. Biol.* **2018**, *53*, 1–28.

(3) Wolynes, P. G.; Eaton, W. A.; Fersht, A. R. Chemical physics of protein folding. *PNAS* **2012**, *109*, 17770–17771.

(4) Englander, S. W.; Mayne, L. The nature of protein folding pathways. *PNAS* **2014**, *111*, 15873–15880.

(5) Karplus, M. Behind the folding funnel diagram. *Nature Chem. Biol.* **2011**, *7*, 401–404.

(6) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338*, 1042–1046.

(7) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *PNAS* **2012**, *109*, 17845–17850.

(8) Orozco, M. A theoretical view of protein dynamics. *Chem. Soc. Rev.* **2014**, *43*, 5051–5066.

(9) Meier, K.; Choutko, A.; Dolenc, J.; Eichenberger, A. P.; Riniker, S.; van Gunsteren, W. F. Multi-Resolution Simulation of Biomolecular Systems: A Review of Methodological Issues. *Angewandte Chem. Int. Ed.* **2013**, *52*, 2820–2834.

(10) Best, R. B. Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol.* **2012**, *22*, 52–61.

(11) Rao, V. V. H. G.; Gosavi, S. Using the folding landscapes of proteins to understand protein function. *Curr. Opin. Struct. Biol.* **2016**, *36*, 67–74.

(12) Dragan, A. I.; Read, C. M.; Crane-Robinson, C. Enthalpy-entropy compensation: the role of solvation. *Eur. Biophys. J.* **2017**, *46*, 301–308.

(13) Roux, B. The calculation of the potential of mean force using computer simulations. *Comp. Phys. Comm.* **1995**, *91*, 275–282.

(14) Meirovitch, H.; Cheluvaraja, S.; White, R. P. Methods for Calculating the Entropy and Free Energy and their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Prot. & Pept. Sci.* **2009**, *10*, 229–243.

(15) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(16) Plotnikov, N. V.; Kamerlin, S. C. L.; Warshel, A. Paradynamics: An Effective and Reliable Model for Ab Initio QM/MM Free-Energy Calculations and Related Tasks. *J. Phys. Chem. B* **2011**, *115*, 7950–7962.

(17) Onuchic, J.; Wolynes, P.; Lutheyschulten, Z.; Socci, N. Toward an outline of the topography of a realistic protein-folding funnel. *PNAS* **1995**, *92*, 3626–3630.

(18) Gershenson, A.; Gierasch, L. M.; Pastore, A.; Radford, S. E. Energy landscapes of functional proteins are inherently risky. *Nature Chem. Biol.* **2014**, *10*, 884–891.

(19) Lin, M. M.; Shorokhov, D.; Zewail, A. H. Dominance of misfolded intermediates in the dynamics of alpha-helix folding. *PNAS* **2014**, *111*, 14424–14429.

(20) Xiong, K.; Asciutto, E. K.; Madura, J. D.; Asher, S. A. Salt Dependence of an alpha-Helical Peptide Folding Energy Landscapes. *Biochemistry* **2009**, *48*, 10818–10826.

(21) Hazel, A.; Chipot, C.; Gumbart, J. C. Thermodynamics of Deca-alanine Folding in Water. *J. Chem. Theor. Comp.* **2014**, *10*, 2836–2844.

(22) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J.Chem. Theory and Computation* **2008**, *4*, 435–447.

(23) Pronk, S.; Pll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(24) MacKerell, A. D. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(25) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.* **2004**, *25*, 1400–1415.

(26) Garcia, A.; Sanbonmatsu, K. alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds. *PNAS* **2002**, *99*, 2782–2787.

(27) Kaminski, G.; Friesner, R.; Tirado-Rives, J.; Jorgensen, W. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(28) Lopez, M.; Chin, D.; Baldwin, R.; Makhatadze, G. The enthalpy of the alanine peptide helix measured by isothermal titration calorimetry using metal-binding to induce helix formation. *PNAS* **2002**, *99*, 1298–1302.

(29) Goch, G.; Maciejczyk, M.; Oleszczuk, M.; Stachowiak, D.; Malicka, J.; Bierzynski, A. Experimental investigation of initial steps of helix propagation in model peptides. *Biochemistry* **2003**, *42*, 6840–6847.

(30) Fitzgerald, J. E.; Jha, A. K.; Sosnick, T. R.; Freed, K. F. Polypeptide motions are dominated by peptide group oscillations resulting from dihedral angle correlations between nearest neighbors. *Biochemistry* **2007**, *46*, 669–682.

(31) Echeverria, I.; Makarov, D. E.; Papoian, G. A. Concerted Dihedral Rotations Give Rise to Internal Friction in Unfolded Proteins. *J. Am. Chem. Soc.* **2014**, *136*, 8708–8713.

(32) Avdoshenko, S. M.; Das, A.; Satija, R.; Papoian, G. A.; Makarov, D. E. Theoretical and computational validation of the Kuhn barrier friction mechanism in unfolded proteins. *Scientific Reports* **2017**, *7*.

(33) Avbelj, F.; Baldwin, R. Limited validity of group additivity for the folding energetics of the peptide group. *PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS* **2006**, *63*, 283–289.

(34) Sheu, S.; Yang, D.; Selzle, H.; Schlag, E. Energetics of hydrogen bonds in peptides. *PNAS* **2003**, *100*, 12683–12687.

(35) Silva, R.; Kubelka, J.; Bour, P.; Decatur, S.; Keiderling, T. Site-specific conformational

determination in thermal unfolding studies of helical peptides using vibrational circular dichroism with isotopic substitution. *PNAS* **2000**, *97*, 8318–8323.

(36) Decatur, S. IR spectroscopy of isotope-labeled helical peptides: Probing the effect of N-acetylation on helix stability. *Biopolymers* **2000**, *54*, 180–185.

(37) Ianoul, A.; Mikhonin, A.; Lednev, I.; Asher, S. UV resonance Raman study of the spatial dependence of alpha-helix unfolding. *J. Phys. Chem. A* **2002**, *106*, 3621–3624.

Acknowledgements

This work was partially supported by the Spanish Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER, UE) under Project CTQ2016-79345-P and by the Fundación Séneca under Project 20789/PI/18.

JC acknowledges a fellowship provided by "Fundación Séneca – Agencia de Ciencia y Tecnología de la Región de Murcia" under the "Saavedra-Fajardo" program (20028/SF/16). The authors would like to thank Prof. Sebastián Fernández-Alberti for constructive criticism of the manuscript. We thank the computational assistance provided by J. F. Hidalgo and F. García.

Author contributions

A.B., J.Z., A.R. and J.C. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

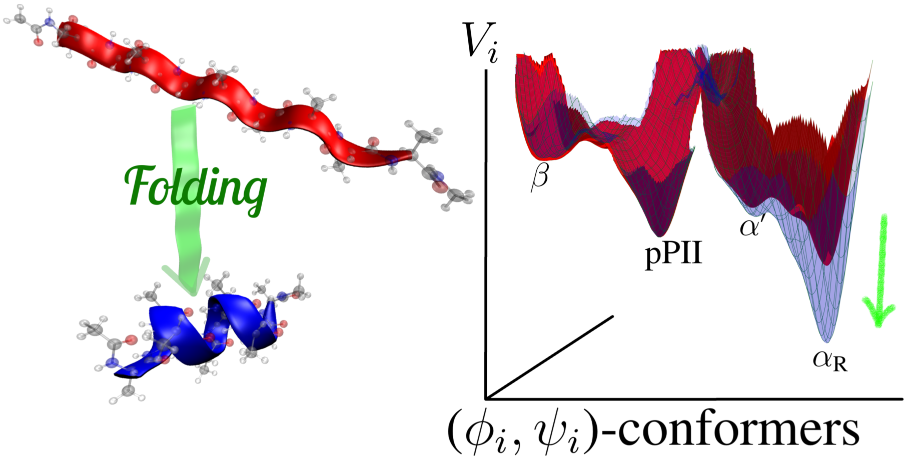
Folding
V_i
β
pPII
α'
α_R
(ϕ_i, ψ_i)-conformers

Figure TOC