Accurate Calculation of Barnase and SNase Folding Energetics Using Short MD Simulations and an Atomistic Model of the Unfolded Ensemble. Evaluation of Force Fields and Water Models

Juan J. Galano-Frutos[†,‡], *Javier Sancho**[†,‡,‖]

[†] Department of Biochemistry and Molecular and Cellular Biology, University of Zaragoza, Pedro Cerbuna 12, Zaragoza 50009, Spain.

[‡] Biocomputation and Complex Systems Physics Institute (Joint Units: BIFI-IQFR (CSIC) and GBsC-CSIC), University of Zaragoza, Mariano Esquillor s/n, Edificio I + D, Zaragoza 50018, Spain.

[‖] Aragon Health Research Institute (IIS Aragón), Avda. San Juan Bosco 13, Zaragoza 50009, Spain.

* jsancho@unizar.es

ABSTRACT: As proteins perform most cellular functions, quantitative understanding of protein energetics is required to gain control of biological phenomena. Accurate models of native proteins can be obtained experimentally but the lack of equally fine models of unfolded ensembles impedes the calculation of protein folding energetics from first principles. Here we show that an atomistic unfolded ensemble model, consisting on a few dozen conformations built from a protein sequence, can be used in conjunction with an X-ray structure of its native state to calculate accurately by difference the changes in enthalpy and in heat capacity of the polypeptide upon folding. The calculation is done using Molecular Dynamics simulations and popular force fields and water models and, for the two model proteins studied (barnase and SNase), the results agree within error or are very close to their experimentally determined properties. The enthalpy sampling of the unfolded ensemble is done through short 2-ns simulations that do not significantly modify the representative distribution of R_g of the starting conformations. The impressive accuracy obtained opens the possibility to investigate quantitatively systems or phenomena not amenable to experiment, and paves the way for addressing the calculation of protein conformational stability (i.e. the change in Gibbs energy upon folding), a central goal of Structural Biology. So far, these calculated enthalpy and heat capacity changes, combined with the experimentally determined melting temperatures of the corresponding protein, allow to reproduce the stability curves of both barnase and SNase.

INTRODUCTION

Proteins are the main constituents and agents in cells but the thermodynamic properties that govern protein forms and their interaction capabilities cannot be calculated from first principles yet.[1] This fact greatly limits the quantitative understanding of biological processes at the cellular level, and severely impairs progress in many other fields such as genetic interpretation, *ab initio* structure prediction, drug design or rational-design protein engineering. At the core of the problem lays the fact that the protein properties (e.g. conformational stability, binding affinity) that govern function are equilibrium properties, so that accurate atomistic models of both the initial and final states are essential to carry out such calculations. However, despite the impressive achievements of Structural Biology in revealing native structures of proteins at atomic resolution, the structural details of the other side of the folding equilibrium remains obscure. This has to do with the unsuitability of current structure determination techniques for providing accurate models of the unfolded ensemble of proteins, which precludes any calculation by difference of properties such as stability.

Conformational stability is a key thermodynamic property of proteins, arising from the balance of numerous interactions established between solvent and proteins atoms in both the native state and the unfolded ensemble.[2,3] To be able to calculate the stability of a protein from first principles, the changes in enthalpy and in entropy associated to the folding reaction must be calculated. An accurate calculation of the entropy term (ΔS_{unf}) is not available yet, but the Gibbs–Helmholtz equation[4] (eq. 1) allows to estimate protein stability in a good approximation by replacing the knowledge of the entropy by that of the heat capacity change plus that of the melting temperature:

$$\Delta G(T) = \Delta H_{Tm} \times (1 - T/T_m) - \Delta Cp \times [T_m - T + T \times \ln(T/T_m)] \quad (1)$$

While thermodynamic parameters such as enthalpy and heat capacity changes can be determined experimentally by differential scanning calorimetry[5] their calculation *in silico* has been traditionally considered difficult for two reasons. First, as advanced above, while the folded state is finely represented by experimental structures, what kind of representation of the unfolded ensemble is realistic enough for performing such calculations is far from clear. Second, due to the extensive cancellation of interactions taking place in protein folding,[6] the energy change of the process is orders of magnitude smaller than the actual energies of the initial and final states. As a consequence, highly accurate values for both the folded state and the unfolded ensemble enthalpies and heat capacities must be calculated, which constitutes a challenge for the available force fields used in Molecular Dynamics (MD) simulations. We show here that short, all-atom, MD simulations of barnase and SNase folded structures and unfolded ensembles (consisting of 40 conformations randomly selected from larger ensembles generated with the ProtSA server[7]) provide enough accuracy and enough sampling to allow calculating the folding energetics of these model proteins within experimental error.

METHODS

PDB structures of the folded states. The X-ray structures with the highest resolution available in the Protein Data Bank for *Bacillus amyloliquefaciens* ribonuclease (barnase; PDB ID: 1A2P[8]) and *Staphylococcus aureus* nuclease (SNase; PDB ID: 2SNS[9]) were selected. Two and thirteen residues, respectively, were missing at the N-terminals of barnase and SNase. The missing residues were added, using Chimera,[10] in an extended conformation, as expected for residues located in protein ends and not contributing to the electron density map. The zinc ion and the

water molecules present in the X-ray structure of barnase, as well as a calcium ion and an organic ligand present in the X-ray structure of SNase were removed.

Generation of unfolded ensembles, filtering of very extended conformations and random selection of the unfolded sample. Large ensembles of unfolded conformations of barnase and SNase (2300 structures each) were generated and stored in individual PDB files using ProtSA,[7] a web server that uses the Flexible-Meccano algorithm[11] for backbone-conformation generation and Sccomp[12] to add the side chains. Within each ensemble, the diameters of the unfolded structures followed asymmetric distributions (Figure S1) where around ~10 % of the structures had particularly long diameters. In order to leave out those very elongated structures and so to avoid the need of using very large water boxes for the MD simulations, we established diameter cut-offs of 12 and 15 nm for the unfolded structures of barnase and SNase, respectively (Figure S1). From each filtered distribution, which contained around 2000 structures, forty unfolded structures were randomly selected for simulation.

MD simulations setup. Three replicas of each of forty unfolded structures and ten replicas of the folded structures of both barnase and SNase were prepared and minimized (emtol = 1.0 kJ/mol) using the steepest descent minimization algorithm included in the GROMACS 4.6.7 package.[13] A dodecahedral box was set with diameters of 14 and 17 nm for folded barnasa and SNase, respectively, and of 15 and 18 nm for the unfolded systems of these proteins, respectively. The number of water molecules in the larger unfolded systems was then evened to that in the folded systems by removing the exceeding waters, and immediately after, a short 0.2ns-NPT-simulation step (0 K and 1 atm) was run to adjust the volume of the unfolded systems. Then, 3 ns of MD simulation were run for each structure, which included heating, equilibration and production stages. Systems were heated up to the target temperature using a

ramp with a Berendsen thermostat[14] (7 steps × 50 ps). Equilibration was performed in three subsequent stages. First, an NVT step (150 ps) to introduce the v-rescale thermostat, second, an NPT step (250 ps) to couple pressure to the system by using a Berendsen barostat,[14] and third, a further NPT step (250 ps) to introduce the more realistic Parrinello-Rahman barostat[15] also used in the productive phase (2 ns). Periodic boundary conditions (PBC) were established and the Partial Mesh Ewald (PME) algorithm[16] was used to treat electrostatic interactions. For van der Waals interactions, a cut-off method with a radius of 1 nm and the Potential-shift-Verlet modifier was used. Bond lengths were constrained using the LINCS algorithm,[17] and the time step was set to 2 fs. Charmm22 with CMAP correction (version 2.0)[18], Amber99SB-ILDN[19], and a recent upgrade of this latter force field, A99SB-disp[20], combined with different models of explicit water (see Table S1) are tested.

Determination of the minimum sample size for the 'unfolded state' model. To estimate the minimum number of unfolded structures (minimum sample size) required to calculate ΔH_{unf} with similar or smaller errors than those reported in the experimental determinations (the highest errors reported for barnase and SNase ΔH_{unf} are 30 kJ/mol and 32 kJ/mol, respectively, see Table 1), we have considered that a half of them (15 kJ/mol and 16 kJ/mol, respectively) arises from variation related to the unfolded ensemble. Thus, since the $\langle H_u \rangle$ values fit a Normal distribution (Figure S2), the minimum sample size of the unfolded ensemble can be calculated[21] as:

$$n = (z_{0.975} \times s_M / d)^2 \quad (2)$$

where $z_{0.975}$=1.96 is the value for the standard Normal distribution with a confidence interval of 95 % (two-tailed); d is the maximum allowed error (half of the experimental ΔH_{unf} error); and s_M is the sample standard deviation (forty unfolded structures sampled).

RESULTS AND DISCUSSION

Proteins analysed and workflow. To structurally describe unfolded ensembles of proteins, a method able to generate a full-atom model of the ensemble has been developed.[22,23] This method adds side chains –in non-clashing conformations selected from a library of rotamers– to backbone conformations of the proper length generated with the distribution of dihedral angles and radius of gyration found in fully unfolded proteins. The method has been implemented in a web application (ProtSA)[7] that, from an input amino acid sequence, builds large ensembles of unfolded conformations and writes the corresponding coordinate files that can be used to calculate properties of the ensemble. We have used ProtSA here to obtain models of the unfolded ensembles of two well-studied proteins in order to assess the reliability of such models through the calculation of unfolding thermodynamic properties from a statistical analysis of MD simulations.

The first protein analysed, barnase, is a 110-residue, extensively characterized[24-26] extracellular ribonuclease from *Bacillus amyloliquefaciens,* whose equilibrium heat denaturation is reversible, follows a two-state mechanism and leads to a fully unfolded ensemble.[26] Barnase lacks disulfide bonds and cofactors, and its three-dimensional structure (Figure 1) has been determined by X-ray crystallography at 1.5 Å resolution.[8] The second protein analysed is a 149-residue extracellular nuclease from *Staphylococcus aureus* (SNase), whose three-dimensional structure (Figure 1) has also been determined by X-ray crystallography at 1.5 Å resolution.[9] SNase sequence is unrelated to that of barnase (sequence identity of 12.1 %). Barnase[24-26] and SNase[27] unfolding thermodynamics has been investigated in detail by many laboratories and a summary of the experimental values found for their corresponding changes in enthalpy and in heat capacity is presented in Table 1.

Table 1. Experimental energetics of Barnase and SNase unfolding.

Protein	Ref.	ΔH_{unf} (pH, T_m)[a] (kJ/mol) (K)	ΔH_{unf} (T_m)[b] (kJ/mol) (K)	ΔCp_{unf}[c] (kJ/mol·K)
Barnase	25	272±29 (2.2, 298)[d,g]	260±29 (295)	
	25	373±29 (3.2, 313)[d,g]	378±29 (315)	3.6±0.5
	25	486±29 (5.5, 328)[d,g]	510±29 (335)	
	25	448±29 (3.8, 321)[e,g]	355±29 (295) 427±29 (315) 499±29 (335)	3.6±0.5
	24	307±15 (2.2, 295)[d,f]	307±15 (295)	7.2±0.7 (295K)
	24	418±21 (3.4, 316)[d,f]	415±21 (315)	6.1±0.6 (315K)
	24	486±24 (5.5, 328)[d,f]	520±26 (335)	4.9±0.5 (335K)
	24	445±22 (4.0, 322)[e,f]	355±29 (295) 427±29 (315) 499±29 (335)	5.7±0.6 (322K)
	26	345±18 (2.0, 297)[d]	333±17 (295)	
	26	449±22 (3.0, 313)[d]	464±23 (315)	6.2±0.8
	26	546±27 (5.0, 327)[d]	596±30 (335)	
	26	523±26 (4.0, 324)[e]	346±17 (295) 470±24 (315) 594±30 (335)	6.2±0.8
SNase	27	180±18 (4.1, 316)[d]	160±16 (307)	2.5±2.0 (307K)
	27	239±24 (4.5, 319)[d]	248±25 (317)	6.1±2.0 (317K)
	27	301±30 (5.5, 325)[d]	320±32 (327)	9.7±2.0 (327K)
	27	301±30 (5.5, 325)[e]	138±14 (307) 228±23 (317) 318±32 (327)	9.0±2.0 (325K)

[a] Experimental barnase and SNase ΔH_{unf} and ΔCp_{unf} values reported at the temperatures (T_m of the DSC experiment) and pH values indicated. These data are used to obtain, by short extrapolation using $\Delta H(T) = \Delta H(T_m) + \Delta Cp \times (T - T_m)$, the corresponding experimental values (reported in the contiguous column at the right) at the indicated temperatures.

[b] Experimental ΔH_{unf} values at the indicated temperatures. As data for barnase come from three different laboratories, the mean experimental value at 315 K, $\langle \Delta H_{unf} \rangle$ = 427±24 kJ/mol (average of the six values reported), can be used to compare with the values calculated *in silico* at the same temperature, which are reported in Tables 2 and S3.

[c] Experimental ΔCp_{unf} values. Those reported in Ref. 24 and 27 are temperature dependent, as indicated. As data for barnase come from three different laboratories a mean experimental value of ΔCp_{unf}, $\langle \Delta Cp_{unf} \rangle$ = 5.3±0.7 kJ/mol·K, has been obtained from 3.6±0.5,[25] 6.1±0.6[24] and 6.2±0.8[26] kJ/mol·K.

[d] Experimental data reported at T_m values closest to the simulation temperatures (see Table 2).

[e] Experimental data reported at pH values closest to the approximate simulation pHs (~4 for barnase, ~4.7-5.8 for SNase, see Table 2).

[f] Data is reported in Ref. 24 without error. Extrapolation to the simulation temperatures has been done assuming 5 % and 10 % errors in the reported ΔH_{unf} and ΔCp_{unf} values, respectively.

[g] Data is reported in Ref. 25 indicating the errors in ΔH_{unf} were always lower than 29 kJ/mol.

The workflow used for the calculations of barnase and SNase folding energetics is summarized in Figure 1. For each of these proteins, the absolute enthalpies of the unfolded ensemble and of the corresponding folded conformation described by the X-ray structure[8,9] is calculated from MD simulation trajectories run with different models of explicit water using the Charmm22/CMAP[18], Amber99SB-ILDN[19], or A99SB-disp[20] (Table S1) force fields. Subsequently, the change in enthalpy upon unfolding is calculated by difference (subtracting the calculated folded value from the unfolded one), and the change in heat capacity by exploiting the linear dependency of ΔH with temperature. Finally, the calculated ΔH_{unf} and ΔC_{punf} values, together with the experimental midpoint denaturation temperature T_m, are used to calculate ΔG_{unf} with the Gibbs-Helmholtz equation[4] (eq. 1).

Calculation of barnase energetics. We calculate the unfolding enthalpy change, ΔH_{unf}, as the difference between $\langle H_u \rangle$, the absolute enthalpy of the unfolded ensemble, and $\langle H_f \rangle$, the absolute enthalpy of the folded state:

$$\Delta H_{unf} = \langle H_u \rangle - \langle H_f \rangle \quad (3)$$

These absolute enthalpies are calculated by averaging enthalpies obtained for individual replicas of each state (H_u^i or H_f^i). On the other hand, each of those replica enthalpies are obtained as

averages of the enthalpy values displayed by 1000 evenly spaced frames extracted from 2 ns-productive trajectories of the specific replica.

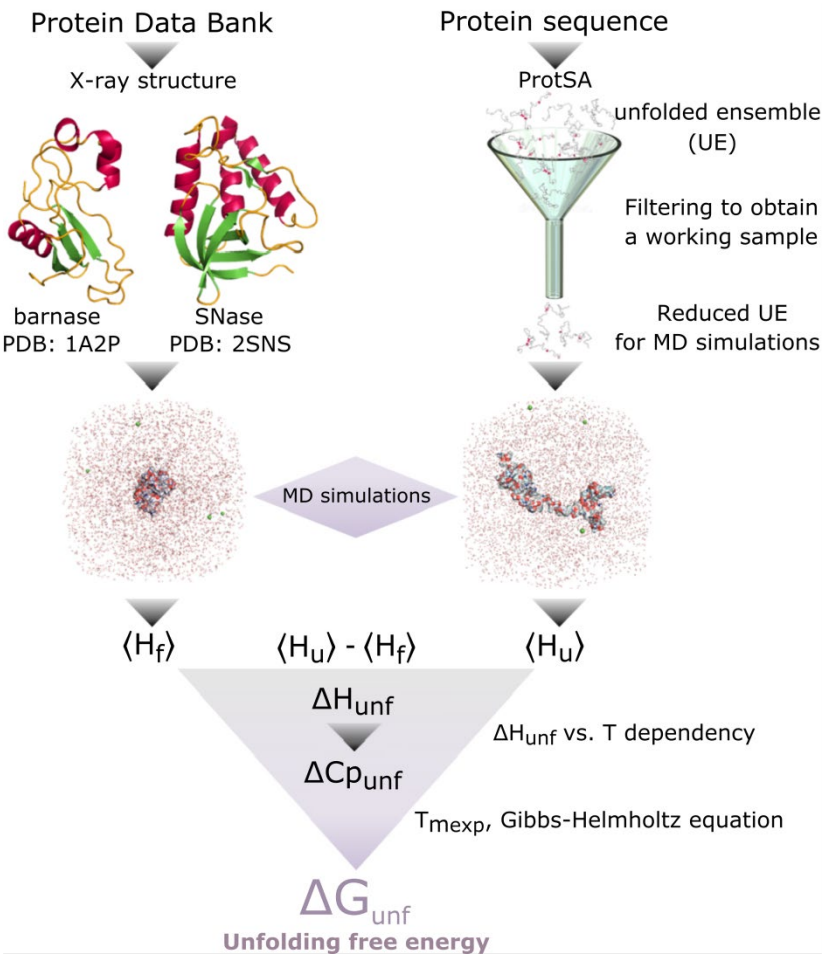

Figure 1. Workflow of the *in silico* approach followed to calculate protein folding energetics. The folded state, represented by the X-ray structure and a reduced sample of an unfolded ensemble generated by the ProtSA server are simulated (full-atom MD simulations) to calculate their averaged absolute enthalpies ($\langle H_f \rangle$ and $\langle H_u \rangle$, see also Figure S2), which are used to calculate ΔH_{unf} at the simulated temperatures, ΔCp_{unf} from the linear ΔH_{unf} vs. T relationship, and ΔG_{unf} by substituting the calculated ΔH_{unf} and ΔCp_{unf} values along with the experimental denaturation temperature T_m in the Gibbs-Helmholtz equation (eq. 1).

Initial simulations for barnase were carried out by combining the three-site model of water Tip3p[28] with the Charmm22/CMAP force field[18] at a simulation temperature of 315 K and a pH of ~ 4 (residue charges: $Asp^-|Glu^0|His^+$) (Tables 2 and S1, setup B0). For the native state, typical

enthalpy values of individual frames mounted to around -2×10^6 kJ/mol. For each trajectory, averaging the enthalpies of 1000 frames provided H_f^i values with standard errors (SE) between 18 and 76 kJ/mol. Importantly, the H_f^i values obtained for the 10 native replicas carried out were similar (higher and lower values differed by 210 kJ/mol, see Figure S2) and their averaging provided for $\langle H_f \rangle$ a value of -2,014,423 ± 21 kJ/mol (mean ± SE). This large $\langle H_f \rangle$ value obtained, which is 3-4 orders of magnitude larger than the experimentally determined ΔH_{unf} (see Table 1), confirmed that calculating ΔH_{unf} by difference is challenging. Nevertheless, we reasoned that such a calculation should be feasible if the small SE obtained for $\langle H_f \rangle$ described the accuracy achieved and if the atomistic model of the unfolded ensemble allowed to calculate $\langle H_u \rangle$ for this ensemble with similar accuracy.

Accordingly, we calculated $\langle H_u \rangle$ from 120 MD trajectories corresponding to 3 replicas of each of 40 different unfolded structures randomly sampled from the barnase unfolded ensemble generated by ProtSA, after having filtered out the highly elongated ones (see Methods). In these simulations, enthalpy values of individual frames of unfolded conformations also mounted to around -2×10^6 kJ/mol and, for each trajectory, averaging the enthalpies of 1000 frames provided H_u^i values with standard errors (SE) between 14 and 90 kJ/mol). The H_u^i values obtained for the 3 replicas of any of the 40 unfolded structures were close to one another, differing in less than 180 kJ/mol. After doing replica averaging, the calculated enthalpies of the 40 structures differed by less than 185 kJ/mol (see Figure S2, setup B0). With these data, the absolute enthalpy for the unfolded state, $\langle H_u \rangle$, was calculated at -2,014,077 ± 6 kJ/mol (mean ± SE) by averaging the 120 H_u^i values. Thus, from equation 2, the enthalpy change of barnase unfolding at the simulation temperature of 315 K resulted in ΔH_{unf} = 346 ± 27 kJ/mol (Table 2, setup B0), in fair agreement with the average experimental value (427 ± 24 kJ/mol, mean ± SE) (see footnote [b] in Table 1).

Once ΔH_{unf} was calculated using setup B0 (the only setup tried at that stage), we determined the sensitivity of the procedure to the force field and the water model used by reproducing the full process using five additional setups (B1-B5), including the use of the Amber99SB-ILDN force field[19] and four additional models of explicit water (Tables 2 and S1). For setups B2-B4 a single-replica approach with the same number of sampled unfolded conformations (40) was used. Results are summarized in Table 2. Compared to Charmm22/CMAP, the Amber99SB-ILDN force field (setup B1) yielded, under otherwise identical simulation conditions, slightly larger (0.08 %) values for folded and unfolded absolute enthalpies, but the ΔH_{unf} value calculated (366 ± 19 kJ/mol) was similar. Compared to Tip3p, the additional three-site models evaluated, Spc[29] (setup B2) and Spc/E[30] (setup B3), as well as the four-site model Tip4p[28] (setup B4) systematically gave lower absolute values of ΔH_{unf} (10, 17 and 46 % lower values at 315 K, respectively) whereas the value obtained using the five-site model Tip5p[31] (setup B5) was 26 % higher: 438 ± 36 kcal/mol, which coincides with the experimental value.

As explicit in eq. 1, in addition to ΔH_{unf}, the change in specific heat capacity that takes place in the folding reaction, ΔCp, is a key thermodynamic magnitude contributing to define the actual stability of a protein at a given temperature. Because ΔCp is the derivative of ΔH with respect to temperature, the sensitivity to temperature of our calculation procedure could be tested by calculating the barnase change in specific heat capacity upon unfolding, ΔCp_{unf}. For the sake of comparison experimental ΔCp_{unf} values[24-26] are given in Table 2. Thus, we calculated the barnase enthalpy change at two additional temperatures (295 and 335 K) in otherwise the same manner. Calculated ΔH_{unf} values showed the expected linear dependency versus temperature (R^2=0.98 for Setup B0; see Figure S3), so that we obtained the barnase ΔCp_{unf} value (ΔCp_{unf} = 3.7 ± 0.4 kJ/mol·K), in reasonable agreement with the average experimental value (5.3 ± 0.7 kJ/mol·K,

mean ± SE, see footnote [c] in Table 1). ΔCp_{unf} estimations with setups B1 and B5 (Table 2 and Figure S3) were also carried out and, as with setup B0, linearity of calculated ΔH_{unf} vs. T was observed. For setup B1 (Amber99SB-ILDN) the calculation of ΔCp_{unf} resulted in 4.5 ± 0.5 kJ/mol·K, whereas for setup B5 (Tip5p) a value of 4.0 ± 0.2 kJ/mol·K was obtained.

Table 2. Calculated energetics of Barnase and SNase unfolding.

Protein	Setup[a]	Setup Code	T (K)	ΔH_{unf} (kJ/mol)	ΔCp_{unf}[b] (kJ/mol·K)
Barnase	Charmm22/CMAP, Tip3p, His^+|Glu^0|Asp^- (pH ~ 4)	B0	295	286±26	3.7±0.4
			315	346±27	
			335	432±21	
	Amber99SB-ILDN, Tip3p, His^+|Glu^0|Asp^- (pH ~ 4)	B1	295	295±14	4.5±0.5
			315	366±19	
			335	473±27	
	Charmm22/CMAP, **Spc**, His^+|Glu^0|Asp^- (pH ~ 4)	B2	315	310±26	-
	Charmm22/CMAP, **Spc/E**, His^+|Glu^0|Asp^- (pH ~ 4)	B3	315	287±25	-
	Charmm22/CMAP, **Tip4p**, His^+|Glu^0|Asp^- (pH ~ 4)	B4	315	187±24	-
	Charmm22/CMAP, **Tip5p**, His^+|Glu^0|Asp^- (pH ~ 4)	B5	295	364±36	4.0±0.2
			315	438±33	
			335	524±33	
SNase	Charmm22/CMAP, Tip3p, His^+|Glu^-|Asp^- (pH ~ 4.7-5.8)	N0	307	195±27	7.4±0.1
			317	268±33	
			327	342±34	
	Charmm22/CMAP, **Tip5p**, His^+|Glu^-|Asp^- (pH ~ 4.7-5.8)	N1	307	321±53	9.6±0.3
			317	411±40	
			327	512±46	

[a] Combinations of force field and water model used in short 2ns-MD simulations of barnase and SNase at each of the temperatures (T) indicated in the corresponding rows, with the protonation states of histidine, glutamic and aspartic acid residues, and the pH modelled. To facilitate comparison of the different setups, the changes introduced relative to the initial setup, B0, are indicated in bold.

[b] ΔCp_{unf} calculated as the slope of a linear representation of calculated ΔH_{unf} vs. simulation temperature. Corresponding fittings for setups and statistical parameters are shown in Figure S3.

Calculation of SNase energetics. To discard that the fine agreement observed between calculated and experimental barnase thermodynamic properties was coincidental, a second protein, SNase, was analysed once all barnase analyses had been completed and the calculation procedure had been fully established. To represent the SNase unfolded ensemble, 40 structures were randomly selected from those generated by ProtSA in the same manner as previously done for barnase (see Methods). The SNase unfolding thermodynamic parameters were determined using Charmm22/CMAP and two alternative water models (see Table 2). Using Tip3p (setup N0), the values calculated (ΔH_{unf} = 268 ± 33 kJ/mol at 317 K, and ΔCp_{unf} = 7.4 ± 0.1 kJ/mol·K) agreed, within error, with the experimental values of 248 ± 25 kJ/mol and 6.1 ± 2.0 kJ/mol·K (Table 1) determined by differential scanning calorimetry.[27] As observed for barnase calculations, the ΔH_{unf} and ΔCp_{unf} calculated using Tip5p (setup N1, one replica for each of 40 unfolded conformations) were larger (411 ± 40 kJ/mol and 9.6 ± 0.3 kJ/mol·K, respectively) that those obtained with Tip3p. Thus, for barnase, the Tip5p (setup B5) model approximated the experimental values slightly better than Tip3p, but for SNase Tip3p (setup N0) was clearly better. The results of the two proteins taken together indicate that the simpler Tip3p model (implemented in setups B0 and N0) can be used for the simulations.

Calculation of excess Cp profiles (DSC thermograms) of barnase and SNase. Calculation of ΔG_{unf} at fixed temperatures. The calculated ΔH_{unf} and ΔCp_{unf} values allowed to obtain[32] the barnase and SNase excess Cp profiles for setups B0 and N0. As shown in Figure 2, these Cp profiles are in good agreement with the corresponding experimental ones. Besides, the *in silico* ΔH_{unf} and ΔCp_{unf} values, in combination with experimental T_m values, allowed to calculate ΔG_{unf} at fixed temperatures using the Gibbs-Helmholtz equation (eq. 1). With this approach, we calculated the conformational stability of barnase at 298 K and that of SNase at 293 K using the

ΔH_{unf} and ΔCp_{unf} values obtained for setups B0, B1, B5, N0 and N1 (Table 2). The conformational stabilities calculated using the data obtained with the Tip3p water model and Charmm22/CMAP (setups B0 and N0) were 27.0 ± 3.5 and 19.1 ± 5.3 kJ/mol, respectively (Table 3), which match within error the experimental values (31.9 ± 2.0 kJ/mol for barnase[24-26, 33-35] and 23.6 ± 3.9 kJ/mol for SNase.[27,36-38] Using the same water model and Amber99SB-ILDN (setup B1), the conformational stability calculated for barnase (28.1 ± 3.6 kJ/mol) was also in excellent agreement with the experimental value. On the other hand, the conformational stabilities calculated for barnase and SNase using the Tip5p water model and Charmm27 (setups B5 and N1) were, respectively, very close (33.4 ± 4.2) or a bit higher in the case of SNase (31.2 ± 8.0) than the experimental values, but still in agreement within the error. Once again, it seems that the simpler Tip3p model may be a good choice to perform these MD-based calculations.

Table 3. Calculated *vs.* experimentally determined barnase and SNase unfolding Gibbs energy changes.

					Calculated								Experimental	
Protein	**Ref.**	**pH**	**T_m[a,b] (K)**	**Setup Code**	**ΔH_{unf}[a,c] (kJ/mol)**	**ΔCp_{unf}[a,d] (kJ/mol·K)**	**ΔG_{unf}[e] (kJ/mol)**	**Ref.**	**Method**	**pH**	**T (K)**	**ΔG_{unf}[f] (kJ/mol)**	**$\langle\Delta G_{unf}\rangle$[g] (kJ/mol)**
Barnase	26	4.0	323.5±0.3	B0	386±24	3.7±0.4	27.0±3.5	34	Spectroscopy	4.0	298	33.9[h]	
				B1	416±20	4.5±0.5	28.1±3.6	24-26,35	DSC	4.0	298	31.8±3.8[i]	31.9±2.0
				B5	476±34	4.0±0.2	33.4±4.2	33	LEM (urea)	4.4	298	29.9±0.2	
SNase	27	5.0	323.8±0.3	N0	318±34	7.4±0.1	19.1±5.3	36	Spectroscopy	7.0	293	24.1±1.3	
								27,38	DSC	7.0	293	21.2±6.4[i]	23.6±3.9
				N1	480±46	9.6±0.3	31.2±8.0	37	LEM (urea)	7.0	293	25.5[h]	

[a] Values used to calculate ΔG_{unf} using the Gibbs-Helmholtz equation (eq. 1).
[b] Experimental T_m reported at pH values close to those used in setups B0 and N0 (Table 2).
[c] Calculated ΔH_{unf}. The linear fits in Figure S3 for the indicated setup have been used to extrapolate the ΔH_{unf} values at the experimental T_m.
[d] Calculated ΔCp_{unf} for the setups indicated (see Table 2 and Figure S3).
[e] ΔG_{unf} values calculated at the temperatures (298 K for barnase and 293 K for SNase) for which experimental values are shown in the Table. Shown errors result from error propagation.
[f] Experimentally determined ΔG_{unf} for barnase and SNase using three different experimental approaches: thermal unfolding followed spectroscopically, thermal unfolding followed calorimetrically (DSC), and chemical denaturation (with urea) analysed with the Linear Extrapolation Method (LEM).
[g] Mean value obtained by averaging the experimental ΔG_{unf} values reported in the indicated references.
[h] No error reported.
[i] Averages obtained from the ΔG_{unf} values reported or calculated (equation 1) using data from the indicated references. ΔG_{unf} values used in averaging were: 29.0 ± 3.7[25], 34.8 ± 4.9[26], 27.5 ± 4.0[24], 35.8 ± 0.2[35], 19.7 ± 8.3[27], and 22.6 ± 4.4[38] kJ/mol.

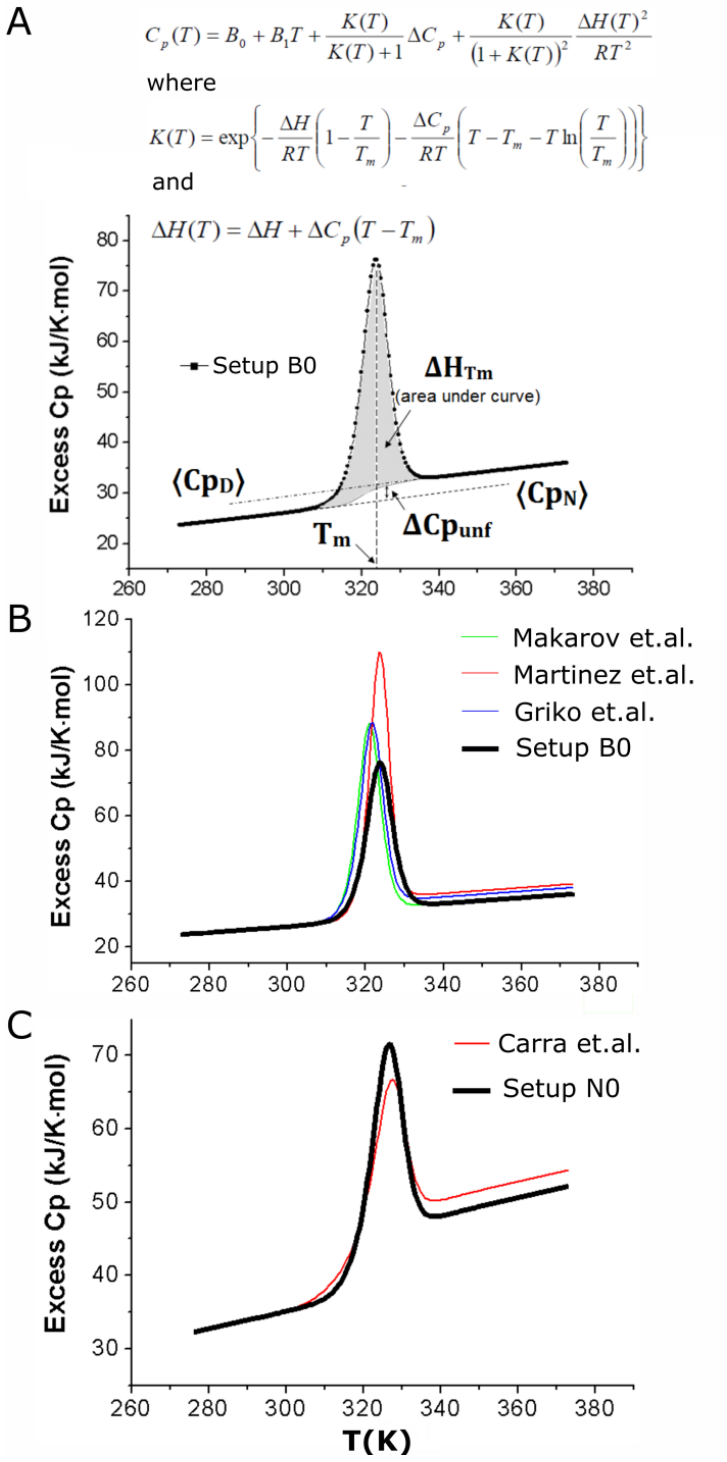

Figure 2. Calculated unfolding thermograms (excess Cp as a function of T). A) Barnase excess Cp profile obtained from the *in silico* calculated ΔH_{unf} and ΔCp_{unf} values for setup B0 (Table 2) and the experimental T_m (323.5 K at pH=4)[26] using the theoretical equations shown.[32] The main characteristic elements of the curve, T_m: the temperature at which 50 % of the protein molecules are unfolded, ΔH_{Tm}: the area under the peak, and ΔCp_{unf}: the molar heat capacity

change of the transition (assumed as constant) are indicated. B) Comparison of the calculated profile displayed in A) with the experimental barnase profiles corresponding to the ΔH_{Tm}, ΔCp_{unf} and T_m values reported in the indicated references (see Table 1). C) Comparison of the SNase calculated (ΔH_{Tm}, ΔCp_{unf} from setup N0, plus experimental T_m=323.8 K at pH=5)[27] and experimental profiles.

Calculation of Gibbs energy profiles (stability curves) for barnase and SNase. The variation of the conformational stability of a protein (i.e. the Gibbs energy difference between the folded and unfolded states) with temperature depends on the enthalpy and heat capacity differences between the two states and on the melting temperature, as is described by the Gibbs–Helmholtz equation (eq. 1). The representation of ΔG_{unf} versus temperature is known as the stability curve of the protein,[4] which is crucial to define the temperature interval where the native protein is conformationally stable and can be functional. In view of the fair agreement between the calculated and the experimental ΔG_{unf} at the specified temperatures, stability curves were calculated for barnase and SNase over the 0-100 ºC temperature interval using the ΔH_{unf} and ΔCp_{unf} values obtained from setups B0 and N0, respectively. The calculated stability curves (Figure 3) match reasonably well the experimental curves, drawn with experimental ΔH_{unf} and ΔCp_{unf} values reported for pH conditions similar to those in setups B0 and N0 (Figure 3). Thus, it appears that using our procedure, the conformational stability of a protein can be calculated in the 0-100 ºC temperature interval in fair agreement with experiment.

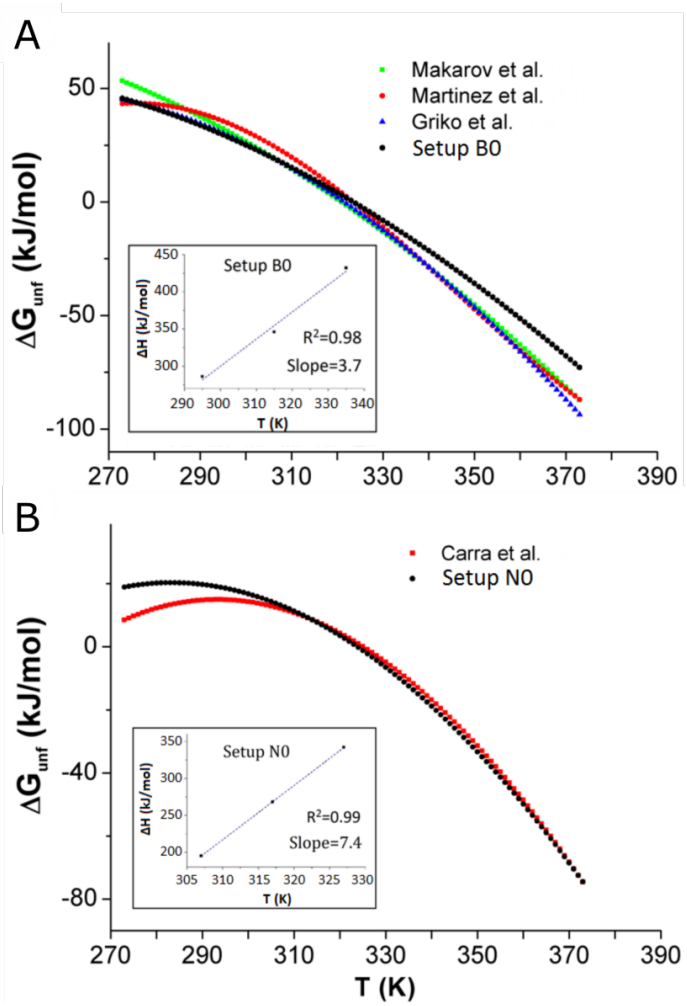

Figure 3. Simulated stability curves: Gibbs energy difference as a function of temperature. Stability curves simulated with eq. 1 using the calculated ΔH_{unf} and ΔCp_{unf} values plus the experimental T_m for (A) barnase (setup B0) and (B) SNase (setup N0), compared to the experimental stability curves. The ΔH_{unf}, ΔCp_{unf} and T_m values used to obtain the simulated Gibbs energy profiles are those displayed in Table 3 (left-hand part), whereas the values taken for the experimental profiles are: Makarov's profile (ΔH_{Tm}=448 kJ/mol, ΔCp=3.6 kJ/mol·K and T_m=320.9 K),[25] Martinez's profile (ΔH_{Tm}=523 kJ/mol, ΔCp=0.6+0.108×T-0.00028×T2 kJ/mol·K and T_m=323.5 K),[26] Griko's profile (ΔH_{Tm}=445 kJ/mol, ΔCp=5.7 kJ/mol·K and T_m=321.5 K)[24] and Carra's profile (ΔH_{Tm}=301 kJ/mol, ΔCp=9.0 kJ/mol·K 1 and T_m=325.1 K).[27] All these experimental values corresponded to unfolding experiments performed in pH conditions similar to those settled in setups B5 and N0. Insets illustrate the linear dependency of the calculated ΔH_{unf} values with T, and the calculated ΔCp_{unf} values (slopes) (see also Figure S3). The Pearson coefficient of the fits, R^2, are shown.

Sample size and computation time required. The above calculation of accurate ΔH_{unf} and ΔCp_{unf} values for barnase (or for SNase) required simulating 1.2 μs, mostly used to simulate different conformations of the unfolded ensemble to provide efficient sampling. The simulations using Tip5p waters took about twice the time as those using Tip3p (Figure S4). No differences were observed between the distributions of average RMSD for the simulated unfolded conformations using either water model (Figure S5).

The sample size required to compute ΔH_{unf} with a given error was determined as explained in Methods. For barnase and SNase (110 and 149-residue-long, respectively), ΔH_{unf} could be calculated with errors lower than those in the experimental data (up to 30 kJ/mol for barnase[24-26] or 32 kJ/mol for SNase;[27] see Table 1) from small samples containing 36 and 42 unfolded structures, respectively (see Table S2). Assuming the same standard errors reported in Table 1, eq. 2 indicates that using samples of half that size (i.e. 18 and 21 unfolded structures, respectively) would increase the errors in the calculated ΔH_{unf} values to 43 and 45 kJ/mol, respectively. Using only one unfolded structure just would not work. According to our data for both the barnase and SNase ensembles, the difference between the higher and lower $\langle H_u \rangle$ values of the individual unfolded conformations simulated is of around 200 kJ/mol for the settings using Tip3p waters and of around 400 kJ/mol for those based on Tip5p ones (not shown). Therefore, accurate ΔH_{unf} values cannot be obtained, with the short 2-ns sampling time used here, representing the unfolded ensemble by means of a single conformation.

The small unfolded ensemble sample sizes required to calculate the energetics of barnase and SNase within error, together with the short simulation time needed for the individual trajectories are encouraging, as they seem to indicate that simulating the energetics of larger proteins may not be computationally expensive. Nevertheless, application of this procedure to much larger

proteins should begin by determining the required sample size of the unfolded ensemble in order to anticipate the computational cost.

It is important to stress that because the conformations of the unfolded ensemble are generated to reproduce experimental parameters (R_g and dihedral angles) corresponding to fully unfolded proteins, the calculation of protein energetics using this model should be restricted to proteins that do not display residual structure when they are thermally unfolded.

Alternatives for conformational sampling of the unfolded ensemble in enthalpy change calculations. In principle, a single unfolded conformation simulated for a long time would also allow for an efficient sampling of the unfolded space and, therefore, for a similarly efficient calculation of the average enthalpy of the unfolded ensemble. Fully extended unfolded models are sometimes used to compute the effect of point mutations of protein energetics, but such extended models would be computationally inefficient as starting conformations for an MD approach due to the size of the water box needed to compute their energies. Besides, not matching typical R_g values and distributions of dihedral angles, they would represent an unrealistic starting point that might require longer simulation times in order to sample the unfolded space well. More important than that, even the simulation of unfolded conformations with realistic starting R_g values poses a problem that has to be addressed. Commonly used force fields combined with popular water models (e.g. Tip3p, Tip4p) have been shown to provide a poor solvation of the unfolded polypeptide, leading to its progressive compaction to R_g values close to those of folded conformations[39-42]. To tackle this problem, new water models have been developed, e.g. Tip4p-d[39] or its further refinement Tip4p-d-mod[20], that improve the solvation of disordered proteins by fixing deficiencies in the modeling of water dispersion interactions. On the other hand, new versions of the Charmm[43] and Amber[20,42,44] force fields have also been

recently developed to try to overcome such a drawback. For instance, a modified version of A99SB-ILDN, A99SB-disp, in conjunction with Tip4p-d-mod was shown to reproduce much better experimental SAXS and NMR data of unfolded conformations while simultaneously providing accurate descriptions of folded proteins[20]. Whether these new force fields and water models can capture accurately the subtle energetics of the folded reaction or still need some further tuning has not been investigated in detail. On the other hand, the extent to which the energies of unfolded conformations compacted along MD simulations differ from the average energy of an unfolded ensemble is also unclear. In the following sections, we make a briefly exploration of these issues.

Extent of R_g reduction in short (2-ns) simulations and its impact in enthalpy calculation. Reduction of R_g in longer (100-ns) simulations. The reduction of the R_g of the unfolded conformations of barnase at the end of the short 2ns-simulations performed is small. Both Charmm22/CMAP[18] and Amber99SB-ILDN[19] force fields, when combined with either Tip3p[28] or Tip5p[31] waters (setups B0, B1 and B5, Tables 2 and S1), only reduce the R_g of barnase unfolded conformations by 10-20 % on average, the greater reductions taking place at higher temperatures (Figures 4 and S6). No statistically significant reduction of the R_g of the native conformations takes place (Figures 4 and S6). Whether R_g reductions of such a small extent modify on average the enthalpy values of the individual unfolded conformations is very unlikely. As Figure S7 shows, there appears to be no correlation between the enthalpy value and the R_g of the individual unfolded conformations of barnase (or of SNase) within the wide range of R_g displayed by the simulated unfolded conformations (1,66-4,47 nm for barnase, and 1,92-5,49 nm for SNase). As it seems, the 2ns-simulations performed very likely preserve the energy of the ensemble despite the 10-20 % reduction in R_g that takes place.

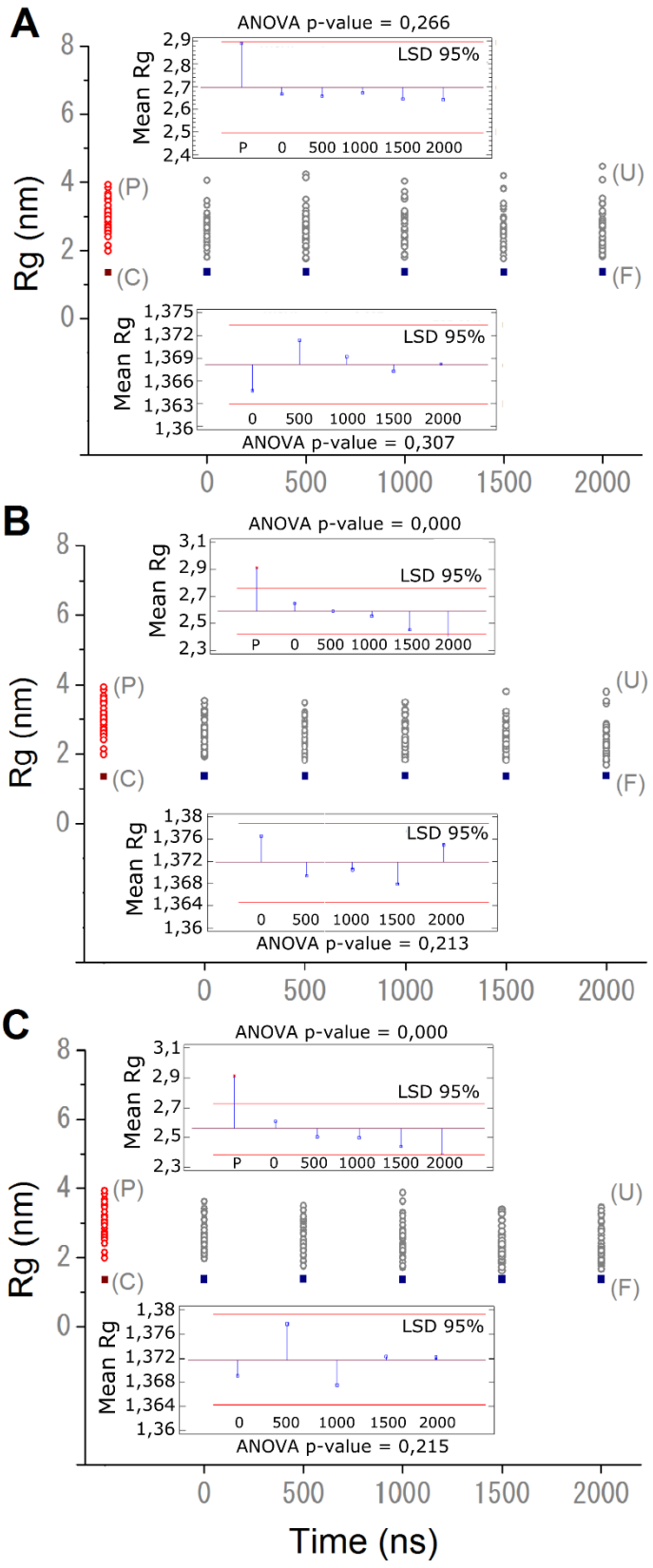
A
ANOVA p-value = 0,266
Mean Rg
LSD 95%
P 0 500 1000 1500 2000
Rg (nm)
(P)
(C)
(U)
(F)
Mean Rg
LSD 95%
0 500 1000 1500 2000
ANOVA p-value = 0,307
0 500 1000 1500 2000
B
ANOVA p-value = 0,000
Mean Rg
LSD 95%
P 0 500 1000 1500 2000
Rg (nm)
(P)
(C)
(U)
(F)
Mean Rg
LSD 95%
0 500 1000 1500 2000
ANOVA p-value = 0,213
0 500 1000 1500 2000
C
ANOVA p-value = 0,000
Mean Rg
LSD 95%
P 0 500 1000 1500 2000
Rg (nm)
(P)
(C)
(U)
(F)
Mean Rg
LSD 95%
0 500 1000 1500 2000
ANOVA p-value = 0,215
0 500 1000 1500 2000
Time (ns)

Figure 4. Evolution of R_g distributions and statistical analysis along 2ns-simulations for setup B0. R_g average distributions of barnase at different times along the simulations for the 40 unfolded (U) structures simulated (black empty circles) and the 10 folded (F) replicas (dark blue filled squares). Charts A, B and C depict distributions and analysis of simulations at the temperatures 295, 315 and 335 K, respectively. For the sake of comparison, red empty circles represent the R_g distribution of the same 40 unfolded structures, as they were provided by ProtSA[7], before being prepared for MD production (labeled as P), and the wine-color square is the R_g value of the crystal structure of barnase (labeled as C). Insets depict a Least Significant Differences (LSD) analysis for unfolded (top) and folded (bottom) R_g distributions, and further include the p-value of an ANOVA mean comparative test.

In fact, if the barnase unfolded ensemble used for the calculation of ΔH_{unf} (40 conformations) is divided into three groups (13 structures with R_g between 1,85 and 2,51, 14 structures with R_g between 2,52 and 2,93, and 13 structures with R_g between 2,94 and 3,79, see Figure S8), the corresponding ΔH_{unf} calculated for each R_g range agree within error (Figure 5) with the global mean obtained for the 40 conformations. The same applies to the SNase ensemble (Figure 5).

To add to previous studies dealing with the evolution of the Rg of unfolded conformations in different water models, we have performed a few longer simulations of barnase at 315 K (3 unfolded structures, 3 replicas each, 100ns-production for each trajectory) using setups B0 and B1 (Charmm22/CMAP or Amber99SB-ILDN+Tip3p), B6 (A99SB-disp+Tip4p-d-mod)[20], and B7 and B8 (Charmm22/CMAP or Amber99SB-ILDN+Tip4p-d[39], respectively). A full description of setups B6-B10 is provided in Table S1. In agreement with previous reports on other systems[39,40], our 100ns-simulations of unfolded barnase in Tip3p strongly reduce R_g from initial values of 1,92, 2,62 or 2,88 nm to 1,5-1,7 nm (Figure S9), close to the R_g of the folded, compact structure (1,4 nm). In contrast, and also in agreement with previous reports[20,39], the simulations of the same unfolded barnase structures with either Tip4p-d or Tip4p-d-mod do not lead to compact conformations (Figure S9).

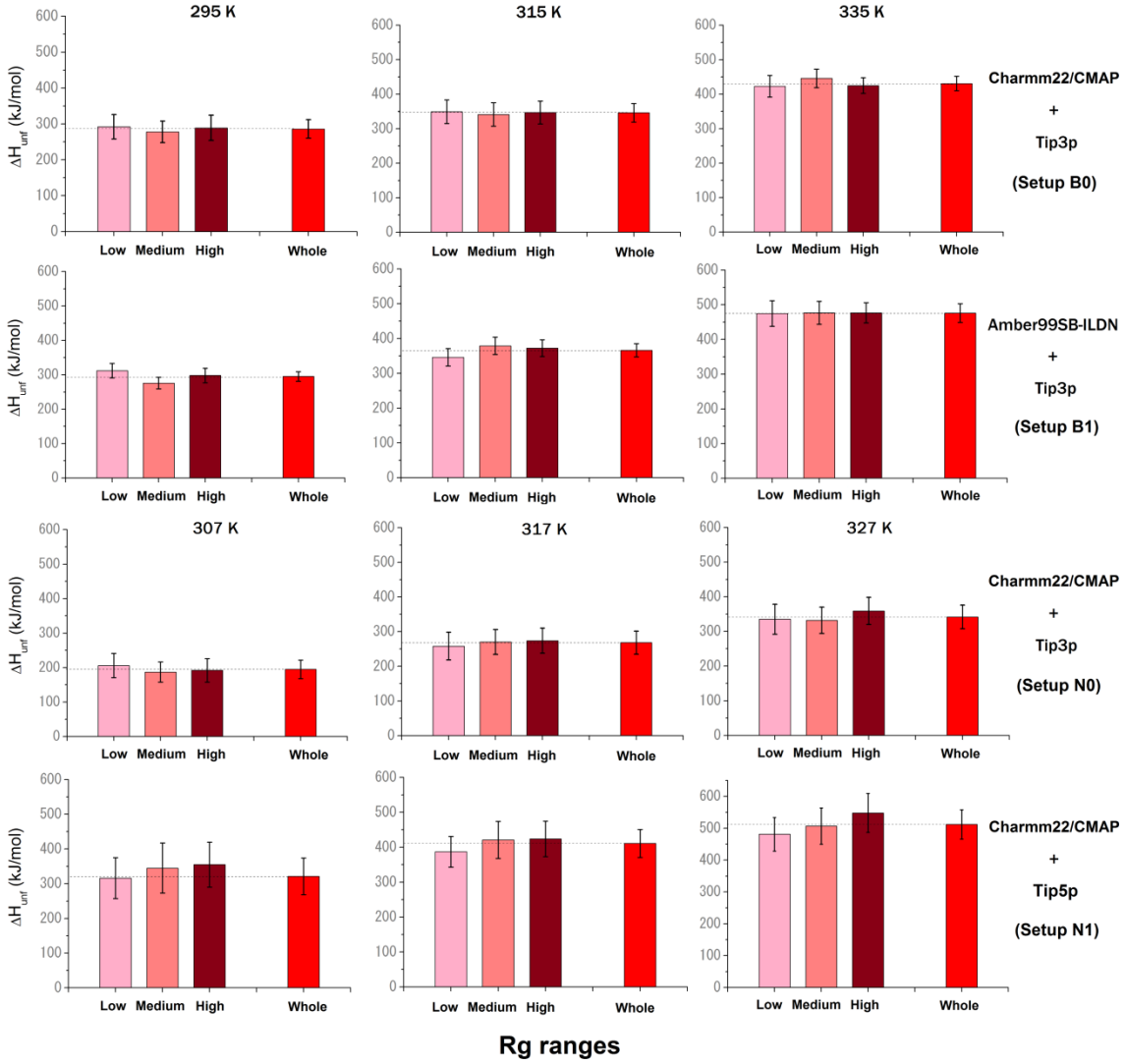

Figure 5. Enthalpy calculation by Rg range of the unfolded ensemble. Calculation of ΔH_{unf} values obtained by partitioning the 40 unfolded structures sampled for barnase and SNase in three sub-sets along the whole range of Rg values (see Figure S8). Low and High ranges of Rg included 13 structures. The Medium range included 14. The ΔH_{unf} values obtained using the whole unfolded ensemble sampled (red bars at the right-hand sides) are included for the sake of comparison. Error bars are indicated, as well as the temperatures and setups used in the simulations (by column and row of charts, respectively).

Need of further tuning of force fields and/or water models that reproduce the R_g of unfolded conformations for their use in calculation of folding enthalpy changes. As the Tip4p-d water model is still in the process of being refined −Tip4p-d-mod represents a try in this direction[20]− it may be useful to assess its performance in energy calculations involving both compact and unfolded structures. Thus, using short 2ns-simulations of the same 40 unfolded structures previously simulated with setups B0-B5 (Tables 2 and S1), we have computed ΔH_{unf} for barnase with setups B6 (A99SB-disp+Tip4p-d-mod), B9 (A99SB-disp+Tip3p) and B10 (Charmm22/CMAP+Tip4p-d-mod) (see R_g profiles in Figure S10 for folded and unfolded structures). In contrast with the results obtained for the water models previously analyzed (described in Table 2), and also in contrast with the experimental values (Table 1), the unfolding enthalpies calculated display negative or very small positive values (setups B6 and B10), or very high positive values doubling the experimental ones (setup B9) (not shown). Also, we notice that, with all these new setups, the enthalpies are calculated with larger standard errors than those calculated with setups described in Table 2. With the values obtained from setups B6, B9 and B10, calculation of ΔC_p from linear regression of enthalpies at different temperatures is not possible. As the Tip4p-d-mod used in setups B6 and B10 has been developed to provide a better solvation of extended protein conformations[20,39] it is possible that it lowers their absolute enthalpies to the point of inverting the sign of the unfolding enthalpy change. Conversely, as A99SB-disp strengthens backbone oxygen/backbone hydrogen interactions[20], which are more abundant in the folded state, it is possible that it lowers the enthalpy of the folded state to a greater extent than that of the unfolded conformations, thus explaining the very high positive values obtained using setup B9 compared to those obtained with setup B1. Perhaps the calculation of folding enthalpy changes from simulation of folded and unfolded structures of

natively structured proteins, such as barnase or SNase, may help further fine tune Tip4p-d and other water models or force fields so that they can reproduce accurately both the geometry of the different protein conformations and the subtle energy changes of the folding reaction. Indeed, using the calculation of folding enthalpies in force-field optimization has been suggested before[40].

Should many short simulations on different starting conformations or a few longer simulations be used, at present, for accurate enthalpy change calculation? Finally, we have explored the possibility of using long simulations of a few unfolded structures as an alternative to our approach based on the sampling of short simulations of a larger number of unfolded conformations. To that end we have used three replicas of each of three barnase unfolded conformations of different R_g (selected from the 40 ones used in the short 2ns-simulations), which were simulated at 315 K using setups B0, B1 and setups B6, B7 and B8 (Table S1). As previously seen for short 2ns-simulations, setup B6 (A99SB-disp+Tip4p-d-mod)[20] provides negative, unphysical, values of ΔH_{unf}, and setups combining Tip4p-d with either CHARMM22/CMAP (B7) or Amber99SB-ILDN (B8) also lead to similar negative results (Table S3). However, setups including Tip3p with either CHARMM22/CMAP (B0) or Amber99SB-ILDN (B1) lead to reasonable values of 244±20 or 218±7 kJ/mol, respectively (Table S3). We notice that these values are, nevertheless, in worse agreement with the experimentally determined one (427±24 kcal/mol, Table 1) that those obtained using the same setups (B0 or B1) from the averaging of short 2ns-simulations (346±27 and 366±19, respectively, see Table 2). Since the simulation time averaged from the long 100-ns-simulations (3×3×100ns=900ns) is larger than that averaged from the short 2ns-ones (40×3×2ns=240ns), it appears that the enthalpy of the unfolded ensemble is better sampled using short simulations of

many different starting conformations than using long simulations of one or a few starting conformations. Certainly, this could change in the future when new force fields/water models that simultaneously capture with high accuracy the geometry and the energetics of disordered conformations become available.

CONCLUSIONS

A simple atomistic model of the unfolded ensemble of proteins, in combination with existing force fields and water models, allows to accurately calculate ΔH_{unf} and ΔCp_{unf} values by difference and to accurately describe the conformational stability of proteins as a function of temperature through the Gibbs-Helmholtz equation (eq. 1). This model of the unfolded ensemble, that circumvent the sampling problem associated to the reduction of R_g in long MD simulations, provides a tool needed[45] for the quantitative understanding of protein folding energetics from first principles. The model may be of help to achieve a more accurate prediction of phenotypes associated to genetic variations, to better understand the contribution of the different elementary interactions to protein stability, or to further fine tune force fields and water models.

ASSOCIATED CONTENT

Supporting Information.

The Supporting Information is available free of charge on the ACS Publications website at DOI:

- Figures S1-S10 and Table S1-S3 (PDF)
- Barnase_unfolded_ensemble_sample (ZIP)
- SNase_unfolded_ensemble_sample (ZIP)

AUTHOR INFORMATION

ORCID

Juan J. Galano-Frutos: 0000-0002-1896-7805

Javier Sancho: 0000-0002-2879-9200

Corresponding Author

*Contact information to whom correspondence should be addressed:

Javier Sancho, E-mail: jsancho@unizar.es, Tel: +34 976 761 286.

Author Contributions

J.S. conceived and directed the investigation. J.J.G-F. carried out and analysed the Molecular Dynamics simulations. J.J.G-F. and J.S. analysed data and wrote the manuscript.

Funding Sources

This work was supported by grants BFU2016-78232-P (MINECO, Spain) and E45_17R (Gobierno de Aragón, Spain).

Notes

The authors declare no conflict of interest.

ACKNOWLEDGMENT

We thank Dr. J. Estrada for explanations about the ProtSA server, and Dr. D. de Sancho and members of the BIFI Institute for scientific discussions. We thank the Biocomputation and Complex Systems Physics Institute (BIFI) of the University of Zaragoza and the Red Española de Supercomputación (RES) for computing facilities granted to perform Molecular Dynamics simulations.

ABBREVIATIONS

Differential Scanning Calorimetry, DSC; LEM, Linear Extrapolation Method; MD, Molecular Dynamics; NMR, Nuclear Magnetic Resonance; R_g, Radius of Gyration; SAXS, Small-Angle X-ray Scattering; SNase, *Staphylococcus aureus* nuclease

REFERENCES

1. Dill, K. A.; MacCallum, J. L., The Protein-Folding Problem, 50 years on. *Science* **2012**, 338, 1042-1046.
2. Deller, M. C.; Kong, L.; Rupp, B., Protein Stability: a Crystallographer's Perspective. *Acta Crystallogr. Sect. F: Struct. Biol. Commun.* **2016**, 72, 72-95.
3. Sancho, J., The Stability of 2-State, 3-State and More-State Proteins from Simple Spectroscopic Techniques... Plus the Structure of the Equilibrium Intermediates at the Same Time. *Arch. Biochem. Biophys.* **2013**, 531, 4-13.
4. Becktel, W. J.; Schellman, J. A., Protein Stability Curves. *Biopolymers* **1987**, 26, 1859-1877.
5. Privalov, P. L.; Dragan, A. I., Microcalorimetry of Biological Macromolecules. *Biophysical Chemistry* **2007**, 126, 16-24.
6. Ben-Naim, A., The Rise and Fall of the Hydrophobic Effect in Protein Folding and Protein-Protein Association and Molecular Recognition. *Open Journal of Biophysics* **2011**, 1, 1-7.
7. Estrada, J.; Bernado, P.; Blackledge, M.; Sancho, J., ProtSA: A Web Application for Calculating Sequence Specific Protein Solvent Accessibilities in the Unfolded Ensemble. *BMC Bioinformatics* **2009**, 10, 104.
8. Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y., Refinement and Structural Analysis of Barnase at 1.5 A Resolution. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **1999**, 55, 386-398.
9. Cotton, F. A.; Hazen, E. E., Jr.; Legg, M. J., Staphylococcal Nuclease: Proposed Mechanism of Action Based on Structure of Enzyme-Thymidine 3',5'-Bisphosphate-

Calcium Ion Complex at 1.5-A Resolution. *Proc. Natl. Acad. Sci. U. S. A.* **1979**, 76, 2551-2555.

10. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, 25, 1605-1612.
11. Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J. R.; Jensen, M. R.; Segard, S.; Bernado, P.; Charavay, C.; Blackledge, M., Flexible-Meccano: A Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and their Associated Experimental Observables. *Bioinformatics* **2012**, 28, 1463-1470.
12. Eyal, E.; Najmanovich, R.; McConkey, B. J.; Edelman, M.; Sobolev, V., Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins. *J. Comput. Chem.* **2004**, 25, 712-724.
13. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, 26, 1701–1718.
14. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, 81, 3684-3690.
15. Parrinello, M.; Rahman, A., Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics* **1981**, 52, 7182-7190.
16. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, 103, 8577-8593.
17. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M., LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, 18, 1463-1472.

18. MacKerell, A. D. J.; 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. Comput. Chem.* **2004**, 25, 1400-1415.

19. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, 78, 1950-1958.

20. Robustelli, P.; Piana, S.; Shaw, D. E., Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proceedings of the National Academy of Sciences* **2018**, 115, E4758.

21. Meeker, W. Q.; Hahn, G. J.; Escobar, L. A., *Statistical Intervals: A Guide for Practitioners and Researchers*. 2nd ed.; John Wiley & Sons: 2017.

22. Bernado, P.; Blackledge, M.; Sancho, J., Sequence-Specific Solvent Accessibilities of Protein Residues in Unfolded Protein Ensembles. *Biophys J* **2006**, 91, 4536-4543.

23. Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M., A Structural Model for Unfolded Proteins from Residual Dipolar Couplings and Small-Angle X-Ray Scattering. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, 102, 17002-17007.

24. Griko, Y. V.; Makhatadze, G. I.; Privalov, P. L.; Hartley, R. W., Thermodynamics of Barnase Unfolding. *Protein Sci.* **1994**, 3, 669-676.

25. Makarov, A. A.; Protasevich, I. I.; Kuznetsova, N. V.; Fedorov, B. B.; Korolev, S. V.; Struminskaya, N. K.; Bazhulina, N. P.; Leshchinskaya, I. B.; Hartley, R. W.; Kirpichnikov, M. P.; Yakovlev, G. I.; Esipova, N. G., Comparative Study of Thermostability and

Structure of Close Homologues--Barnase and Binase. *J Biomol Struct Dyn* **1993**, 10, 1047-1065.

26. Martinez, J. C.; el Harrous, M.; Filimonov, V. V.; Mateo, P. L.; Fersht, A. R., A Calorimetric Study of the Thermal Stability of Barnase and its Interaction with 3'GMP. *Biochemistry* **1994**, 33, 3919-26.
27. Carra, J. H.; Anderson, E. A.; Privalov, P. L., Thermodynamics of Staphylococcal Nuclease Denaturation. I. The Acid-Denatured State. *Protein Sci* **1994**, 3, 944-951.
28. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, 79, 926-935.
29. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J., *Intermolecular Forces*. Reidel, Dordrecht, 1981; p 331.
30. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P., The Missing Term in Effective Pair Potentials. *J. Chem. Phys.* **1987**, 91, 6269-6271.
31. Mahoney, M. W.; Jorgensen, W. L., A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, 112, 8910-8922.
32. Mazurenko, S.; Kunka, A.; Beerens, K.; Johnson, C. M.; Damborsky, J.; Prokop, Z., Exploration of Protein Unfolding by Modelling Calorimetry Data from Reheating. *Sci Rep* **2017**, 7, 16321.
33. Matouschek, A.; Matthews, J. M.; Johnson, C. M.; Fersht, A. R., Extrapolation to Water of Kinetic and Equilibrium Data for the Unfolding of Barnase in Urea Solutions. *Protein Engineering, Design and Selection* **1994**, 7, 1089-1095.

34. Oliveberg, M.; Vuilleumier, S.; Fersht, A. R., Thermodynamic Study of the Acid Denaturation of Barnase and its Dependence on Ionic Strength: Evidence for Residual Electrostatic Interactions in the Acid/Thermally Denatured State. *Biochemistry* **1994**, 33, 8826-8832.

35. Vuilleumier, S.; Fersht, A. R., Insertion in Barnase of a Loop Sequence from Ribonuclease T1. *Eur. J. Biochem.* **1994**, 221, 1003-1012.

36. Eftink, M. R.; Ghiron, C. A.; Kautz, R. A.; Fox, R. O., Fluorescence and Conformational Stability Studies of Staphylococcus Nuclease and its Mutants, Including the Less Stable Nuclease-Concanavalin A Hybrids. *Biochemistry* **1991**, 30, 1193-1199.

37. Shortle, D.; Meeker, A. K., Mutant Forms of Staphylococcal Nuclease with Altered Patterns of Guanidine Hydrochloride and Urea Denaturation. *Proteins* **1988**, 1, 81-89.

38. Shortle, D.; Meeker, A. K.; Freire, E., Stability Mutants of Staphylococcal Nuclease: Large Compensating Enthalpy-Entropy Changes for the Reversible Denaturation Reaction. *Biochemistry* **1988**, 27, 4761-4768.

39. Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E., Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B* **2015**, 119, 5113-5123.

40. Piana, S.; Klepeis, J. L.; Shaw, D. E., Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. *Current Opinion in Structural Biology* **2014**, 24, 98-105.

41. Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J., Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *The Journal of Physical Chemistry Letters* **2019**, 10, 2227-2234.

42. Best, R. B.; Zheng, W.; Mittal, J., Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* **2014**, 10, 5113-5124.

43. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D., CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nature Methods* **2016**, 14, 71-73.

44. Nerenberg, P. S.; Jo, B.; So, C.; Tripathy, A.; Head-Gordon, T., Optimizing Solute–Water van der Waals Interactions To Reproduce Solvation Free Energies. *The Journal of Physical Chemistry B* **2012**, 116, 4524-4534.

45. Eichenberger, A. P.; van Gunsteren, W. F.; Riniker, S.; von Ziegler, L.; Hansen, N., The Key to Predicting the Stability of Protein Mutants Lies in an Accurate Description and Proper Configurational Sampling of the Folded and Denatured States. *Biochim. Biophys. Acta* **2015**, 1850, 983-995.

For Table of Contents Only

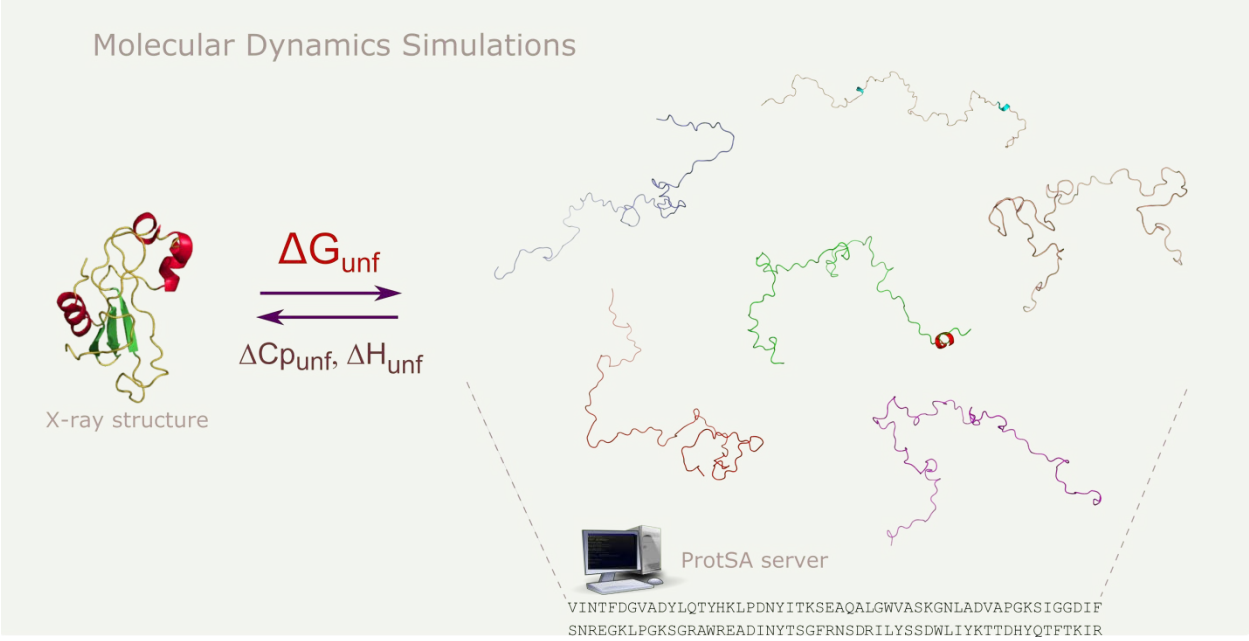

Figure upgraded