SUPPORTING INFORMATION

Accurate Calculation of Barnase and SNase Folding Energetics Using Short Molecular Dynamics 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

Figure S1. Unfolded ensemble diameters distributions. Diameters distributions of the unfolded structures for barnase (top) and nuclease (bottom). Initial distributions obtained from ProtSA are shown in the left panels, A): barnase, C): nuclease, whereas filtered distributions after removing the more elongated structures appear in the right panels, B): barnase, D): nuclease. Insets are included (top right) showing histograms of the diameter distributions along with their typified asymmetry (skewness).

Figure S2. Histograms of absolute enthalpies. Histograms of the enthalpy distributions of the folded (red) and unfolded (blue) replicas (H_f^i and H_u^i , respectively) simulated under conditions specified in setups B0 and N0 (Table 2 of the main text). For the folded state, the enthalpy values of the 10 replicas launched are depicted. For the unfolded state, 40 enthalpies are represented, each an average of the three replicas run for a given unfolded structure. The absolute temperatures used for the simulations are indicated.

Figure S3. Enthalpy change vs. temperature plots. Calculated unfolding enthalpy vs. simulation temperature plots showing linear fits, the determination coefficients (Pearson squared) and slopes ($\Delta C_{p_{unf}}$) for the indicated conditions setups (Table 2 of the main text) under which calculations were carried out.

Figure S4. Simulation cost (time) per water model. Simulation time regarding the water model used in the different setups described in Table 2. Wall-clock measures were obtained from .log files issued by GROMACS during the productive 2ns-phase carried out. In all cases simulations were performed by setting 16 cores to be used in our local computation cluster. The rates shown in the graph were calculated dividing wall-clock values by the number of water molecules (in thousands) included in the simulated systems.

Figure S5. Average RMSD between unfolded and crystal structures. Distributions, for two water models, of the average RMSD between the unfolded conformations simulated for A) barnase and B) SNase (2ns-productive trajectories of each of the 40 conformations simulated) and the corresponding folded crystal structures. Insets display, for each unfolded conformation, the RMSD difference obtained when using Tip3p minus that using Tip5p.

Figure S6. R_g profiles and statistical analysis of short 2ns-simulations for Setup B1. R_g average distributions of barnase at different times along the simulations for the 40 unfolded structures simulated (top part of each combined chart) and the 10 folded replicas (bottom part of each combined chart). For the sake of comparison, green points represent: in bottom-part charts, the R_g of the crystal structure of barnase (labeled as C); in top-part charts, the R_g distribution of the unfolded ensemble sampled from the ProtSA initial distribution (labeled as P). Insets depict a Least Significant Differences (LSD) analysis (95 % confidence intervals), and further include the p-value of an ANOVA mean comparative test. Temperatures used for each chart are indicated.

Figure S7. R_g profiles and statistical analysis of short 2ns-simulations for Setup B5. R_g average distributions of barnase at different times along the simulations for the 40 unfolded structures simulated (top part of each combined chart) and the 10 folded replicas (bottom part of each combined chart). For the sake of comparison, green points represent: in bottom-part charts, the R_g of the crystal structure of barnase (labeled as C); in top-part charts, the R_g distribution of the unfolded ensemble sampled from the ProtSA initial distribution (labeled as P). Insets depict a Least Significant Differences (LSD) analysis (95 % confidence intervals), and further include the p-value of an ANOVA mean comparative test. Temperatures used for each chart are indicated.

Figure S8. Enthalpy *versus* R_g correlation. Scatter plots of individually calculated ΔH_{unf} values for each unfolded structure (2ns-simulations) *versus* their R_g . Each ΔH_{unf} value was calculated by difference using eq. 3 of the main text and the averaged absolute enthalpy for 10 replicas of folded barnase (see the manuscript). Determination coefficients are indicated to show the correlation obtained. Temperatures and setups used in the simulations appear indicated for each column and row of charts, respectively.

Figure S9. R_g distributions and partition of R_g range. R_g distribution of the whole unfolded ensemble modeled by ProtSA (left-hand charts) for A) barnase and B) SNase. Partition of the R_g intervals for the 40 unfolded structures sampled in the 2ns-MD simulations approach are depicted in the right-hand scatter plots. Low, medium and high intervals included 13, 14 and 13 structures, respectively, which were used for the segmented ΔH_{unf} calculation shown in Figure 5.

Figure S10. R_g profiles of unfolded barnase along 100ns-MD simulations. R_g versus time plots along trajectories for indicated setups simulated at 315 K. Three unfolded structures, namely, p355, p1355 and p1955 (three replicas each) are depicted for each setup tested.

Figure S11. Evolution of R_g distributions and statistical analysis along 2-ns simulations for setup B6. R_g average distributions of barnase at different times along the simulations for the 40 unfolded structures simulated (top part of each combined chart) and the 10 folded replicas (bottom part of each combined chart). For the sake of comparison, green points represent: in bottom-part charts, the R_g of the crystal structure of barnase (labeled as C); in top-part charts, the R_g distribution of the unfolded ensemble sampled from the ProtSA initial distribution (labeled as P). Insets depict a Least Significant Differences (LSD) analysis (95 % confidence intervals), and further include the p-value of an ANOVA mean comparative test. Temperatures used for each chart are indicated.

Figure S12. Evolution of R_g distributions and statistical analysis along 2-ns simulations for setup B9. R_g average distributions of barnase at different times along the simulations for the 40 unfolded structures simulated (top part of each combined chart) and the 10 folded replicas (bottom part of each combined chart). For the sake of comparison, green points represent: in bottom-part charts, the R_g of the crystal structure of barnase (labeled as C); in top-part charts, the R_g distribution of the unfolded ensemble sampled from the ProtSA initial distribution (labeled as P). Insets depict a Least Significant Differences (LSD) analysis (95 % confidence intervals), and further include the p-value of an ANOVA mean comparative test. Temperatures used for each chart are indicated.

Figure S13. Evolution of R_g distributions and statistical analysis along 2-ns simulations for setup B10. R_g average distributions of barnase at different times along the simulations for the 40 unfolded structures simulated (top part of each combined chart) and the 10 folded replicas (bottom part of each combined chart). For the sake of comparison, green points represent: in bottom-part charts, the R_g of the crystal structure of barnase (labeled as C); in top-part charts, the R_g distribution of the unfolded ensemble sampled from the ProtSA initial distribution (labeled as P). Insets depict a Least Significant Differences (LSD) analysis (95 % confidence intervals), and further include the p-value of an ANOVA mean comparative test. Temperatures used for each chart are indicated.

Table S1. Summary of force field/water model and conditions setups used for energetics calculations in this work.

Protein	Setup^[a]	Setup Code	T (K)	Simulation Length^[b]
Barnase	Charmm22/CMAP, Tip3p, His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B0	295 315 335	2 ns 100 ns ^[c]
	Amber99SB-ILDN , Tip3p, His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B1	295 315 335	2 ns 100 ns ^[c]
	Charmm22/CMAP, Spc , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B2	315	2 ns
	Charmm22/CMAP, Spc/E , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B3	315	2 ns
	Charmm22/CMAP, Tip4p , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B4	315	2 ns
	Charmm22/CMAP, Tip5p , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B5	295 315 335	2 ns
	A99SB-disp , Tip4p-d-mod , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B6	295 315 335	2 ns 100 ns ^[c]
	Charmm22/CMAP, Tip4p-d , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B7	315	100 ns ^[c]
	Amber99SB-ILDN , Tip4p-d , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B8	315	100 ns ^[c]
	A99SB-disp , Tip3p, His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B9	295 315 335	2 ns
SNase	Charmm22/CMAP, Tip4p-d-mod , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4)	B10	295 315 335	2 ns
	Charmm27, Tip3p, His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4.7-5.8)	N0	307 317 327	2 ns
	Charmm27, Tip5p , His ⁺ Glu ⁰ Asp ⁻ (pH ~ 4.7-5.8)	N1	307 317 327	2 ns

^[a] Combinations of force field and water model, the protonation state of the acid residues: histidine, glutamic and aspartic, and the pH modelled for the carried out simulations of barnase and SNase. To facilitate comparison of the different setups, the changes introduced relative to the initial setup, B0, are indicated in bold.

^[b] The simulation length indicated refers to that run for the unfolded ensemble (40 unfolded structures sampled). In all the cases, the folded state (10 replicas of the native protein) was simulated for 2ns.

^[c] Three unfolded structures randomly taken from the prior sampled 40-conformation ensemble, with three replicas each, were simulated.

Table S2. Minimum sample sizes for the 'unfolded state' model.

Protein	T (K)	$s_M \langle H_u \rangle$^[a] (kJ/mol)	d^[b] (kJ/mol)	Calculated minimum sample size (n)^[c]
Barnase (setup B0)	295	41	15	29
	315	48		39
	335	49		41
		46		36
SNase (setup N0)	297	44	16	29
	307	63		60
	317	49		36
		52		42

^[a] Standard deviations of the absolute enthalpies of forty unfolded structures sampled from the initial unfolded ensemble generated by ProtSA. Averages are shown in bold at the bottom.

^[b] Errors calculated as half of the highest total error reported for the ΔH_{unf} estimations (30 kJ/mol for barnase and 32 kJ/mol for SNase, see also Table 1 of the main text).

^[c] Values obtained from eq. 2 in the Methods section. Averages shown in bold at the bottom.

Table S3. Calculated enthalpies and setups used for long 100ns-simulations of barnase.

Setup code ^[a]	T (K)	Struct. & Rep. ^[b]	$\Delta H_{\text{unf}}^{[c]}$ (kJ/mol)
B0	315	p355-r1	353
		p355-r2	233
		p355-r3	343
		p1355-r1	223
		p1355-r2	233
		p1355-r3	203
		p1955-r1	233
		p1955-r2	183
		p1955-r3	193
			244±20
B1	315	p355-r1	221
		p355-r2	211
		p355-r3	221
		p1355-r1	181
		p1355-r2	241
		p1355-r3	251
		p1955-r1	211
		p1955-r2	221
		p1955-r3	201
			218±7
B6	315	p355-r1	-275
		p355-r2	-145
		p355-r3	-225
		p1355-r1	-195
		p1355-r2	-305
		p1355-r3	-365
		p1955-r1	-415
		p1955-r2	-215
		p1955-r3	-335
			-275±28
B7	315	p355-r1	57
		p355-r2	-113
		p355-r3	-73
		p1355-r1	-143
		p1355-r2	-193
		p1355-r3	97
		p1955-r1	237
		p1955-r2	-193
		p1955-r3	-143
			52±47

B8	315	p355-r1	-64
		p355-r2	-144
		p355-r3	-44
		p1355-r1	46
		p1355-r2	166
		p1355-r3	-334
		p1955-r1	6
		p1955-r2	-104
		p1955-r3	246
		25±57	

^[a] Combinations of force field and water model used, the protonation state of the acid residues: histidine, glutamic and aspartic, and the pH modelled for each setup in Table S1.

^[b] Nine individual replicas of unfolded barnase simulated for each setup, namely, three structures (each with three replicas) randomly selected from those previously sampled in short simulations.

^[c] ΔH_{unf} values obtained by difference using eq. 3 of the main text. The enthalpy of folded barnase is computed from averaging 10 replicas of 2ns-simulations as for the prior calculations. At the end of each setup, a row shows the average with the standard error of the computed ΔH_{unf} .