PCCP

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2014, **16**, 2748

Folding and unfolding thermodynamics of the TC10b Trp-cage miniprotein†

Charles A. English[a] and Angel E. García*[b]

Received 14th October 2013,
Accepted 12th December 2013

DOI: 10.1039/c3cp54339k

www.rsc.org/pccp

We examine the folding–unfolding of a variant of the Trp-cage, known as TC10b, and compare structural stability, dynamics, and thermodynamics with that of the TC5b variant, using replica exchange molecular dynamics (REMD). The TC10b variant was designed to have larger helical stability by the substitution of amino acids with greater alpha helical propensities in the N-terminal region. Experiments have shown TC10b to possess larger overall stability than TC5b. Simulations starting from unbiased, unfolded initial conditions are run for 1 μs per replica. The calculations show a higher melting temperature for TC10b than TC5b, and suggest a more ordered folded structure through the elimination of a substate found in the folded ensemble of TC5b. We model the difference in Gibbs free energy, $\Delta G(P,T)$, of folding using the bootstrap statistical method, which is used to calculate uncertainties associated with the thermodynamic parameters for both variants of the Trp-cage. We find that while the shape of the area for which the protein is stability folded is elliptical for TC5b, there is a degree of uncertainty associated with that of TC10b, with one model suggesting elliptical and another suggesting hyperbolic. This model suggests that at high pressures, TC5b can experience pressure denaturation, but TC10b may not.

1 Introduction

The effects of a small number of mutations on protein structure and stability are of interest in a wide range of protein studies. In particular, proteins such as amyloid beta may have aggregation propensities and properties significantly enhanced or reduced by small mutations on their sequences, such as by the oxidation of the methionine 35 residue.[1] Furthermore, evidence suggests that mutations of a single amino acid, amino acid 22, of the Aβ sequence can drastically impact the behavior of the amyloid.[2–4] Additionally, the study of the effects of sequence mutations leading to divergent evolutionary protein family trees through homologous intermediaries is of vast importance.[5] Comparison of the unique structural features along each branch of the evolutionary family tree for a given protein often involves a small number of variations in protein sequence, therefore determination of the effects of those sequence mutations is of interest. For the reasons given, highly detailed characterizations of the effects of small changes in protein sequence are desirable.

The study of the effects of small changes in protein sequences faces several obstacles. The size of many proteins of interest and the small, perturbative effects of these mutations make experiments difficult and prohibit the use of coarse grain models to study these systems computationally. Furthermore, the nature of the small, perturbative effects necessitates the use of computationally taxing explicit solvent, all-atom models, thus limiting the scope of applications to fast-folding, small model proteins (less than 50 residues), such as the Trp-cage. Finally, computational explorations as to the structural properties of mutational perturbations are also limited to systems for which there already exists much data on the folding process to serve as a point of comparison. The Trp-cage is a model protein which adequately fits these criteria.

The Trp-cage is a small, 20 residue, well-studied miniprotein with an experimental folding time of 4 μs.[6] Originally designed in 2002, the Trp-cage consists of an alpha helix (residues 1–8), a 3–10 turn (residues 12–14), and a polyproline II segment (residues 16–20).[7] The main chain of the protein forms a cage around the hydrophobic tryptophan (W6), shielding it from the solvent. The structure is further stabilized by an aspartic acid arginine salt bridge (D9–R16). The most commonly studied variant, TC5b, has been extensively represented in the literature with studies benchmarking force fields with both implicit and explicit solvents,[8–18] simulation techniques,[9,10,19–24] protein design methods,[25–28] thermodynamic,[29] kinetics,[6,30–32] denaturation,[33] and confinement.[34] A more stable variant known as TC10b was described by Andersen *et al.*[28] The TC10b (PDB code 2JOF) variant has been infrequently studied either experimentally[35] or computationally.[36] Both variants are nearly identical and possess

[a] *Department of Physics, Applied Physics and Astronomy, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. E-mail: englic@rpi.edu*
[b] *Department of Physics, Applied Physics and Astronomy, and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. E-mail: angel@rpi.edu*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3cp54339k

nearly the same folded structure. TC5b, with a sequence of NLYIQWLKDGPSSGRPPPS, possesses a nearly identical sequence to TC10b and the C_α RMSD values for the PDB structures within 0.1 nm of each other. The sequences are different for the first four residues, NIYL in TC5b, are mutated to DAYA in TC10b. Theoretically, the substitution of alanines has been proposed to increase the stability of the alpha helix.[37] It has also been proposed that the overall preference in alpha helicies is slightly more for an aspartic acid than for asparagine in the N-cap region provided the third residue of the helix is tyrosine.[38] Following extensive studies performed on the effects of force fields,[22] water models, confinement,[34] denaturants[39–41] and protonation of side chains[42] on TC5b, the effects of the mutations in the sequence of TC5b and their effect on the stability, structural properties, and thermodynamics of the Trp-cage are characterized.

Using replica exchange molecular dynamics (REMD)[43] the structural ensemble of TC10b Trp-cage was characterized and compared with that of TC5b to determine the effects of small mutations of the overall structural and thermodynamic properties of the Trp-cage. It was found that TC10b was stabilized over TC5b by the mutations. It was also found that the mutations increase the stability of the alpha helix in the folded state alone by calculating the helicity as a function of residue and as a function of temperature. Furthermore, the rigidity of the protein in the folded state was also increased with the elimination of a conformation of the folded state, as found by RMSD histograms and clustering analysis. Furthermore, thermodynamics parameters were calculated using the Hawley equation[44] were found and the associated errors found using bootstrapping.[45] Additionally, *P–T* diagrams were calculated for both TC10b and TC5b, showing at which pressures and temperatures the protein is stably folded. Results give TC5b an elliptical phase diagram and TC10b a hyperbolic phase diagram. The hyperbolic shape, however, is only found to be slightly more likely than an elliptical shape, as found in the estimations of errors.

2 Methods

REMD simulations of both TC10b and TC5b with sequences Ace-NLYIQWLKDGPSSGRPPPS-NMeth and Ace-DAYAQWLKDG PSSGRPPPS-NMeth, respectively, were performed using the GRO MACS 4.5.1 software package.[46] The results of the TC5b simulation have been described in an earlier publication.[47] Following the work in this previous study, which produced thermodynamics parameters that agreed well with experimental data, the choice of forcefield and simulation parameters were kept consistent. The REMD simulation[43] starting from unfolded initial conditions (UIC) was performed in a cubic box solvated with 2863 TIP3P water molecules[48] with side lengths of 4.49 nm. No charged ions were added to the solvent as the TC10b sequence's net charge is zero. This REMD simulation consisted of 50 replicas with a temperature range between 275 K and 559 K with temperatures chosen to ensure a 25% exchange rate using the method outlined in ref. 49. PME electrostatics[50] were used with a 0.12 nm grid spacing and a real space cutoff of 1.0 nm. Van der Waals interactions were also cut off at 1.0 nm. The stochastic dynamics integrator was chosen based upon the findings of Cooke and Schmidler.[51] An extended protein sequence was generated in PyMOL[52] and then simulated in vacuum at 600 K to produce a collapsed unfolded structure. The simulation box was equilibrated at constant volume to 300 K for 10 ps using the Berendsen thermostat[53] and the pressure equilibrated for 20 ps to 1 bar using the Berendsen pressure coupling. The results were verified for equilibration of pressure by extending the constant pressure equilibration simulation to 10 ns and finding the box size to remain practically unchanged. Production REMD simulations were conducted at constant volume. Starting configurations were randomly chosen from the final configurations of thirteen diagnostic simulations run at 25 K temperature intervals from 275 K to 600 K. The energies from these diagnostic simulations were used to calculated replica temperatures which would yield the uniform exchange rate desired. The UIC simulation was run for 1 μs per replica, resulting in an aggregate 50 μs simulation time. A second simulation, using the same parameters and volume, but starting from folded initial conditions (FIC) was run so that convergence of the simulation could be ascertained. This simulation ran for 400 ns per replica with 50 replicas for an aggregate 20 μs simulation time. By this time, the two simulations had converged to the same fraction of folded replicas value, as described below. The last 600 ns of the UIC simulation was used for analysis.

A pressure–temperature stability diagram and the associated thermodynamic parameters was computed with a fitting method based on χ squared analysis. The parameters calculated were: the change in $\Delta\alpha_U(T_o, P_o)$, the coefficient of linear expansion, $\Delta\beta_U(T_o, P_o)$, the isothermal compressibility, $\Delta C_p(T_o, P_o)$, the change in heat capacity at constant pressure, $\Delta S_U(T_o, P_o)$, the change in entropy, $\Delta V_U(T_o, P_o)$, the change in volume, and $\Delta G_U(T_o, P_o)$, the change in free energy, as the protein goes from the folded state to the unfolded state.[54] The free energy difference for the unfolding of a protein at a given temperature and pressure with respect to some reference temperature and pressure is given by the equation:[44]

$$\begin{aligned}\Delta G(T,P) &= \Delta G_U - \Delta S_U(T - T_o) \\ &\quad - \Delta C_p\left[T\left(\log\left(\frac{T}{T_o}\right) - 1\right) + T_o\right] \\ &\quad + \Delta V_U(P - P_o) + \frac{1}{2}\Delta\beta_U(P - P_o)^2 \\ &\quad + \Delta\alpha_U(P - P_o)(T - T_o)\end{aligned} \tag{1}$$

The function to be minimized is:

$$\begin{aligned}\chi^2 &= \sum_i \left[\frac{\Delta G_i - \Delta G(T_i, P_i)}{\sigma_{\Delta G_i}}\right]^2 + \sum_i \left[\frac{x_i^{\text{folded}} - x(T_i, P_i)}{\sigma_{x_i}}\right]^2 \\ &\quad + \sum_i \left[\frac{\Delta E_i - \Delta E(T_i, P_i)}{\sigma_{\Delta E_i}}\right]^2\end{aligned} \tag{2}$$

where $\sigma_{\Delta Y_i}$ is the uncertainty in the average of Y, where Y is the relevant corresponding quantity in the equation (above) for ΔG_U, ΔE, and x the fraction folded, while i is the corresponding replica

number. ΔG and ΔE are the changes in free energy and energy, respectively. The change in ΔG_U can be obtained by the expression

$$\Delta G(P, T) = -k_B T \log\left(\frac{x_u}{x_f}\right) + \langle P \rangle_{T,V} V \qquad (3)$$

where x_u is the fraction of replicas unfolded, x_f the fraction of replicas folded, k_B the Boltzmann constant, T the temperature, V the volume, and P the pressure. Therefore, the value of $\Delta G(P, T)$ in eqn (1) can be computed directly from simulation data with eqn (3), leaving $\Delta\alpha_U(T_o, P_o)$, $\Delta\beta_U(T_o, P_o)$, $\Delta C_p(T_o, P_o)$, $\Delta S_U(T_o, P_o)$, $\Delta V_U(T_o, P_o)$, and $\Delta G_U(T_o, P_o)$, the fit parameters, to be solved for. In terms of simulation quantities, the fitting (performed by eqn (2)) necessitates the calculation of six input values: the fraction of replicas both folded and unfolded, x_f and x_u, respectively, both of which can be computed from the RMSD values with respect to the PDB structure, the average temperature, the average pressure, and the average energy of both folded and unfolded states, sorted into folded or unfolded states by the corresponding RMSD values.

The Newton–Raphson algorithm was chosen as the minimization method, since it was found to quickly converge to the absolute minimum after around five iterations. Comparisons of two data sets required an analysis of the errors on the parameters found. The bootstrap method was chosen to analyze the error associated with the thermodynamic parameters. The bootstrap method is a well-known error determination method which reparameterizes the data set which then is used to check consistency in a computed result.[45] In this work, a set of data containing 10% of the original number of data points was calculated. From these new data sets, thermodynamic parameters were calculated. In total, 300 individual data sets were generated and 300 corresponding sets of thermodynamic parameters were calculated for both TC5b and TC10b. Averages and standard deviations could then be calculated from the distribution of thermodynamic parameters gained from the bootstrap method.

3 Results and discussion

To assess the progress of the folding, the number of replicas folded over the course of the simulation was tracked over time. Additionally, the number of replicas folded were contrasted with the status of the folded initial condition (FIC) simulation. A replica was defined as folded if the C_α RMSD was less than 0.22 nm with respect to the PDB structure (see Fig. 1a). It was then determined based on the steady state nature of the fraction folded graph that 1 μs per replica was sufficient sampling. Convergence was determined by the FIC equilibrating to similar fraction folded values and was determined to have been reached by 400 ns. Therefore, the last 600 ns of the simulation was utilized to compute ensemble averages. As shown in Fig. 1c, the fraction folded *versus* temperature curves are similar for the UIC and FIC simulations. By inspection of the graph of the fraction folded time history correlation function (see Fig. 1b), the correlation time found was 100 ns.[55] Time block units of twice the correlation time, 200 ns, were taken to be the minimum time required for the simulation to display all characteristic behavior of the system and were thus used as a basis to estimate errors.

Verification of the increased stability of the TC10b variant over the TC5b variant, as experiments by Andersen *et al.* suggest,[28] was sought. A melting curve for the protein was constructed, plotting fraction folded *vs.* temperature. The melting temperature was defined as the temperature at which half the frames are folded. By definition, then, the melting temperature

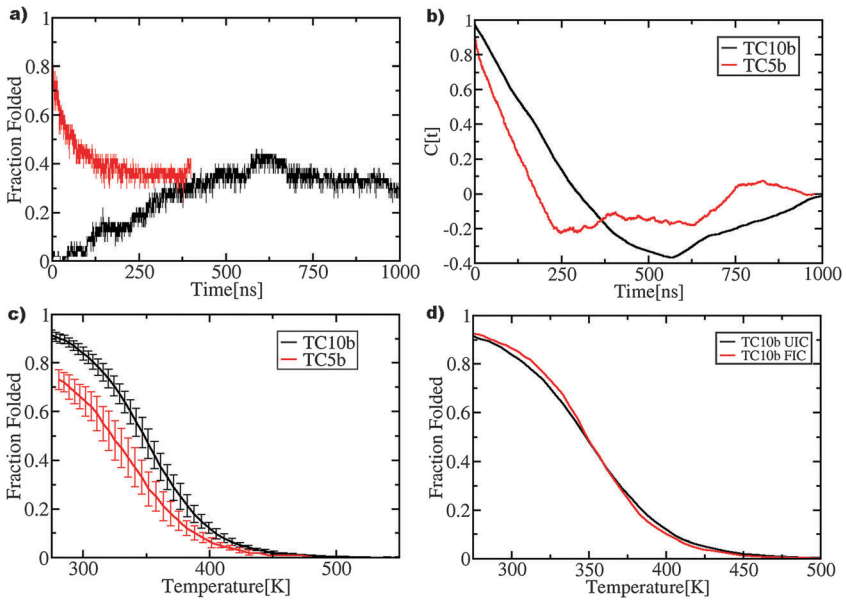

Fig. 1 (a) Measure of the fraction of the 50 replicas folded as a function of time elapsed in the simulation for the FIC and UIC (black and red, respectively). A replica was considered folded if its C_α RMSD was less than 0.22 nm compared to the PDB structure. (b) Correlation functions of the fraction folded time histories for TC10b (black) and TC5b (red). The correlation time for each are 100 ns and 50 ns, respectively. (c) Folding curves for TC10b (black) and TC5b (red) as a function of temperature. (d) Comparison of folding curves as a function of temperature for the TC10b UIC and FIC simulations.

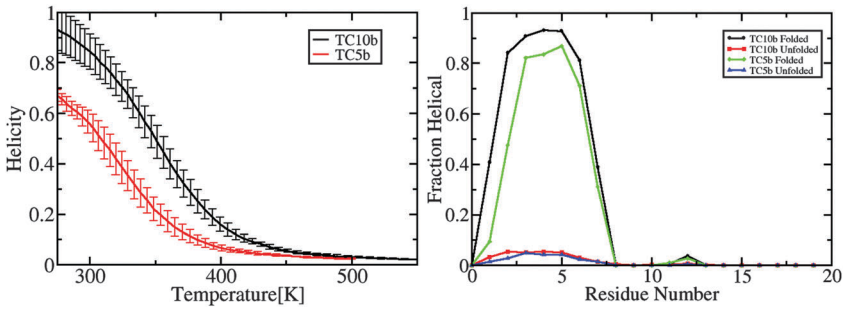

Fig. 2 (left) Measures of the average helicity by temperature of TC10b and TC5b. (right) Measures of the average helicity by residue, subdivided by folding status.

was determined to be 348 ± 8 K. To determine if the mutations did impact the stability of the alpha helix, the helicity *vs.* temperature and the helicity by residue of the protein (Fig. 2) were computed. As in previous work,[47] a protein residue was defined as helical if the following criteria were met: first, the dihedral angles had to correspond to the correct section of the Ramachandran diagram. The accepted range of values was selected to be between −90 and −30 degrees for the ϕ angle and between −17 and −77 degrees for the ψ angle.[56] Secondly, the residue's nearest neighbors were required to have fulfilled the Ramachandran plot criteria as well, in order to ensure the characteristic helical hydrogen bonds had formed. The total number of helical residues was computed for each frame in the production stages of the constant volume ensembles at equilibrium and averaged over the number of frames in the same simulation. The percent of the production simulation in which a given residue was helical was calculated and averaged over the whole simulation to compute the average helicity by residue at 300 K. It was found that TC10b is much more helical as expected, with an average fractional helicity of 0.95 at 300 K *vs.* 0.85 for the same temperature in the TC5b system. Little to no alpha helical content in the 3–10 turn region was found. Helicity gain in the folded state for the first two residues was much greater, by about 50% over TC5b, than for the remain residues associated with the alpha helix, which showed relatively smaller gains of about 10%. Given that the only difference between the two sequences is the replacement of three of the first four residues in the protein, the results discussed so far suggest the increased helical propensity of the alanines is correlated with greater thermal stability of the TC10b Trp-cage miniprotein. In contrast to the folded state, the unfolded state showed no change in overall helicity due to the mutations.

TC5b was found in previous studies to possess a secondary folded substate which consisted a disruption of the N-terminal region of the alpha helix in addition to the shifting of the 3–10 turn. The presence of this second folded substate, as TC5b possesses, was sought for TC10b. The RMSD has been found to accurately describe the folded state of proteins and is able to distinguish the second folded substate for TC5b. For this reason, the C_α RMSD *vs.* radius of gyration histogram at 300 K was chosen to compare the two Trp-cage variants. No distinct second folded substate was detected in the RMSD *vs.* radius of gyration histogram of TC10b. As confirmation that no such substate exists, trajectories were run through the Daura clustering algorithm as provided in the GROMACS software package,[57] employing a cut off of 0.10 nm, and analyzing the lowest temperature trajectory. The top 10 centroid structures, accounting for approximately 90% of the ensemble, were found to correspond to folded states but none of the top centroids included a state with all the features of the secondary folded substate present in TC5b (see Fig. 3).

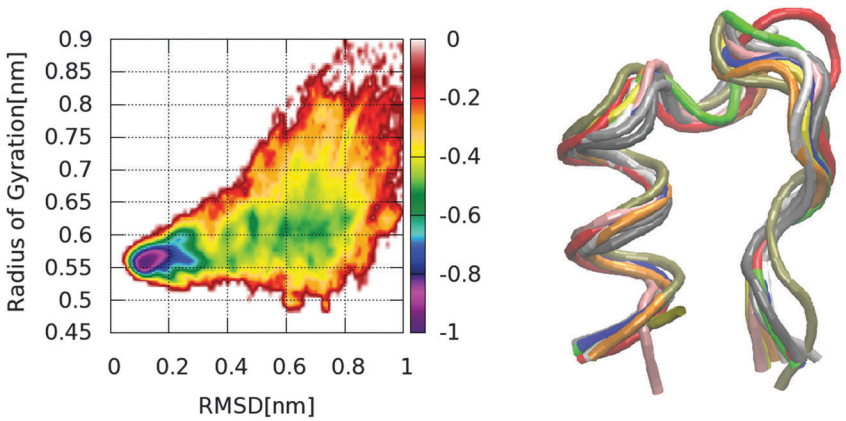

Fig. 3 (left) Histogram of RMSD and radius of gyration. The folded state (violet regions) does not display a secondary substate as in the case of TC5b. (right) Top 10 centroids for TC10b with Daura clustering and a 0.10 nm cut off. While two of the ten centroids show the N-terminal residue disrupted, neither of these centroids possess the 3–10 turn motif associated with the secondary folded substate.

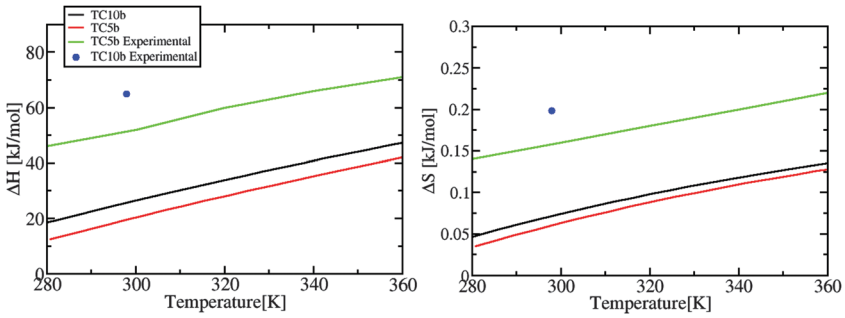

Fig. 4 (left) Enthalpy (right) entropy curves for TC10b (black) and TC5b (red). Included are experimental results from Wafer *et al.*[74] for TC5b (green) and experimental results from Barua *et al.*[28] for TC10b (blue dot).

Table 1 Table of values for thermodynamic parameters at T_o and P_o, 298 K and 0.1 MPa, respectively

Parameter	TC5b	TC10b
$\Delta\alpha(T_o, P_o)$	0.53 ± 0.58 kJ mol^{-1}	0.08 ± 0.83 kJ mol^{-1}
$\Delta\beta(T_o, P_o)$	-1.2 ± 1.7 kJ mol^{-1} MPa^{-2}	-0.02 ± 2.64 kJ mol^{-1} MPa^{-2}
$\Delta C_p(T_o, P_o)$	0.63 ± 0.33 kJ mol^{-1} K^{-1}	0.35 ± 0.51 kJ mol^{-1} K^{-1}
$\Delta S(T_o, P_o)$	0.068 ± 0.031 kJ mol^{-1} K^{-1}	0.068 ± 0.011 kJ mol^{-1} K^{-1}
$\Delta V(T_o, P_o)$	-0.2 ± 20 ml mol^{-1}	-0.2 ± 12 ml mol^{-1}
$\Delta G(T_o, P_o)$	1.01 ± 0.47 kJ mol^{-1}	4.58 ± 0.10 kJ mol^{-1}

A comparison of the thermodynamics of TC10b with those found for TC5b was performed. As a point of comparison, results based on previous work from the data in Day *et al.* for TC5b were re-analyzed.[47] In contrast to the analysis described in the work of Day *et al.*,[47] only the low pressure simulation data were included in the calculations in order to enable a direct comparison between calculated values. High pressures generally denature the protein, as increasing pressure decreases the volume of the protein and thus destabilizes the protein as it requires either a reduction in the size of cavities in the protein, or the forcing of water molecules into the hydrophobic core of the protein,[58,59] or more hydrophobic residues are forced to interact with a greater concentration of water molecules.[22,60–62] The resulting *P–T* diagrams for TC5b and TC10b are given in Fig. 5, with TC5b's yielding an elliptical shape and TC10b yielding a hyperbolic shape. The hyperbolic shape suggests that TC10b is not destabilized by pressure. The final distributions of the thermodynamic parameter distributions are shown in Fig. 6. These results suggested significant differences in $\Delta G_U(T_o, P_o)$, $\Delta V_U(T_o, P_o)$, and $\Delta S_U(T_o, P_o)$. Calculated averages and standard deviations of the fitted parameters are given in Table 1. The entropy and enthalpy as a function of temperature were also computed (see Fig. 4). These calculations suggest higher values at all temperatures considered for TC10b over TC5b, while underestimating entropy and enthalpy values compared to experimental data.

4 Conclusions

The thermodynamic stability of the structure for TC10b was obtained from REMD simulations. The calculated structure was consistent with the PDB structure, with the best structure having an RMSD value of 0.063 nm. The melting temperature was found to be higher than that of TC5b and evidence was presented that the alpha helix is indeed more stable than that of TC5b. Additionally, while the radius of gyration is not substantially different from that of TC5b, the unfolded state of TC10b is only slightly larger than the unfolded state of TC10b. By comparison, the average radius of gyration for all 28 structures in the PDB is 0.72 ± 0.01 nm, matching well with the simulation results for TC10b of 0.68 ± 0.01 nm. The radius of gyration for the unfolded state is not substantially larger than that of the folded state for both TC5b and TC10b. Most of the unfolded ensemble tends to be composed of relatively compact states, except at high temperatures, suggested by the fact that non-native contacts are made between sidechains in the unfolded ensemble.[63] These findings suggesting that the results obtained here are in reasonable agreement with experimental data. The melting temperature calculated was higher than the experimentally obtained result, 348 ± 8 K *versus* 331 ± 2 K,[28] respectively. It has been found that the AMBER99sb forcefield better reproduces the experimentally found NMR order parameter measurements than other force fields in ubiquitin.[64] However, in contrast to work which has shown that the AMBER99sb fails to readily form alpha helices at room temperature for independent alpha helicies,[65] contacts within the tertiary structure of the Trp-cage may work to stabilize the alpha helix. The AMBER99sb forcefield correctly predicted an increase in stability associated with the mutations. As expected and shown by Andersen, alanines proved to stabilize the alpha helix, particularly in the N-terminal region. The data suggested the mutations in TC10b prevented the disruption of the N-terminal region of the alpha helix, thus eliminating the presence of the secondary folded substate possessing this feature.

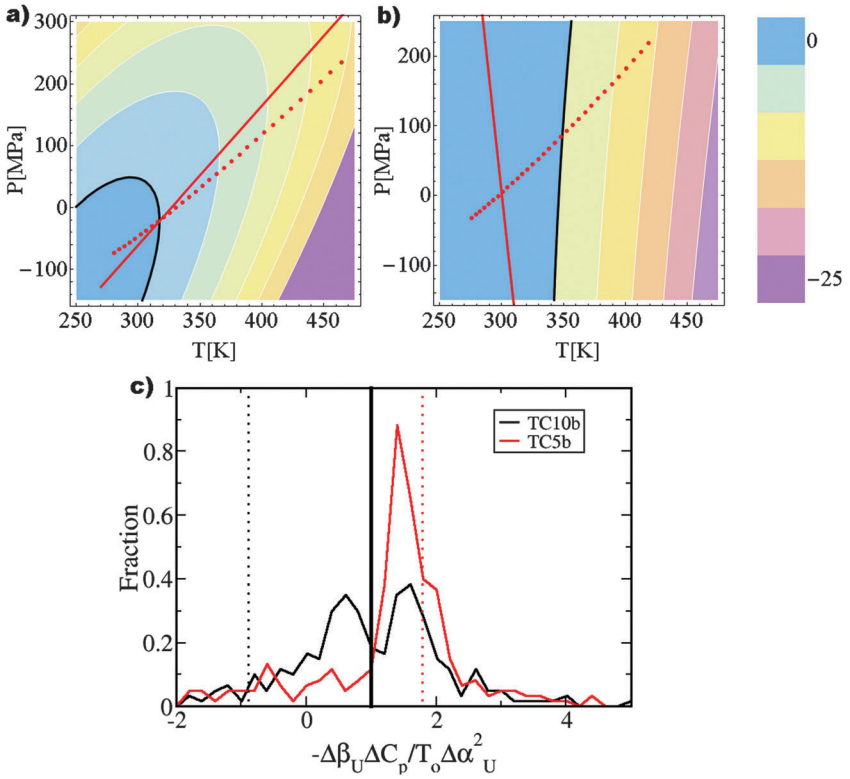

Fig. 5 (top) P–T phase diagrams for TC5b and TC10b, left and right, respectively. The black contour corresponds to zero change in free energy. Dots represents data points for the simulation, the red line corresponds to $\Delta V = 0$. (bottom) Distribution of bootstrap data for elliptic condition, $-\Delta\beta_U\Delta C_p/T_o\Delta\alpha_U^2$. All values of less than one yields hyperbolic phase diagrams, greater than one, elliptical. The dotted lines correspond to the parameters calculated from the whole data set.

An elliptical P–T diagram was calculated for TC5b, representing both denaturation at high and low temperatures and pressures. In contrast, a hyperbolic P–T diagram was calculated for TC10b. An elliptical phase diagram has been commonly observed for globular proteins.[66–68] A P–T diagram is elliptical if the condition

$$\Delta\alpha^2 < \frac{-\Delta C_p \Delta\beta_U}{T_o} \qquad (4)$$

is satisfied.[44] To estimate the error associated with the shape of the phase diagram, the 300 bootstrap parameters were used to compute the right hand side quantity. The resulting histograms are shown in Fig. 6. While the overwhelming majority of the data for TC5b satisfies the condition for an elliptical phase diagram, TC10b gives an ambiguous result, with data concentrated into two peaks, one in the elliptical region, the other in the hyperbolic region. Three thermodynamic parameters associated with the simulation were not significantly different for TC5b and TC10b (*i.e.* $\Delta\alpha_U(T_o, P_o)$, $\Delta\beta_U(T_o, P_o)$, $\Delta C_p(T_o, P_o)$), while the other three, $\Delta S_U(T_o, P_o)$, $\Delta V_U(T_o, P_o)$, and $\Delta G_U(T_o, P_o)$ upon unfolding were different. The larger change in free energy was expected and is due mainly to the increase in stability of the protein. The larger change in entropy observed for TC10b may be explained by the apparent lack of a secondary folded substate in TC10b, implying TC10b lacks the wider range of possible conformational dynamics present for the folded TC5b state and is therefore more ordered than TC5b. This conclusion was also hinted at by the stabilizing role the mutations play in the alpha helix, particularly the N-terminal region, the lack of a secondary substate, and the additional folded stability. These findings suggested that the folded state of TC10b has lower configurational entropy, relative to the entropy of the corresponding unfolded states than the folded state of TC5b, hence the change upon going from the folded state to unfolded state is greater in the case of TC10b than for TC5b. Finally, while the change in volume is approximately zero ml mol^{-1} for TC10b, for TC5b it is non-zero. This finding may be explained by the elimination of the secondary folded substate, limiting the overall range of motion available to the protein in the folded structural ensemble due to a more compact alpha helix. In general, alpha helicies are relatively compact, are highly packed, and are thus not destabilized under high hydrostatic pressure.[69,70] We also note that the calculated relative entropy and relative enthalpy for TC10b, while both higher than that at all temperatures than for TC5b, still remain lower than the experimental values for TC5b.[29,47]

The results showed that small number of mutations can have significant effects on the folded structural ensemble of a protein, in particular the range of conformational dynamics possible, for instance by giving a folded protein more order in its overall structure. By employing error analysis through the bootstrap method, the effects of these mutations on the thermodynamics

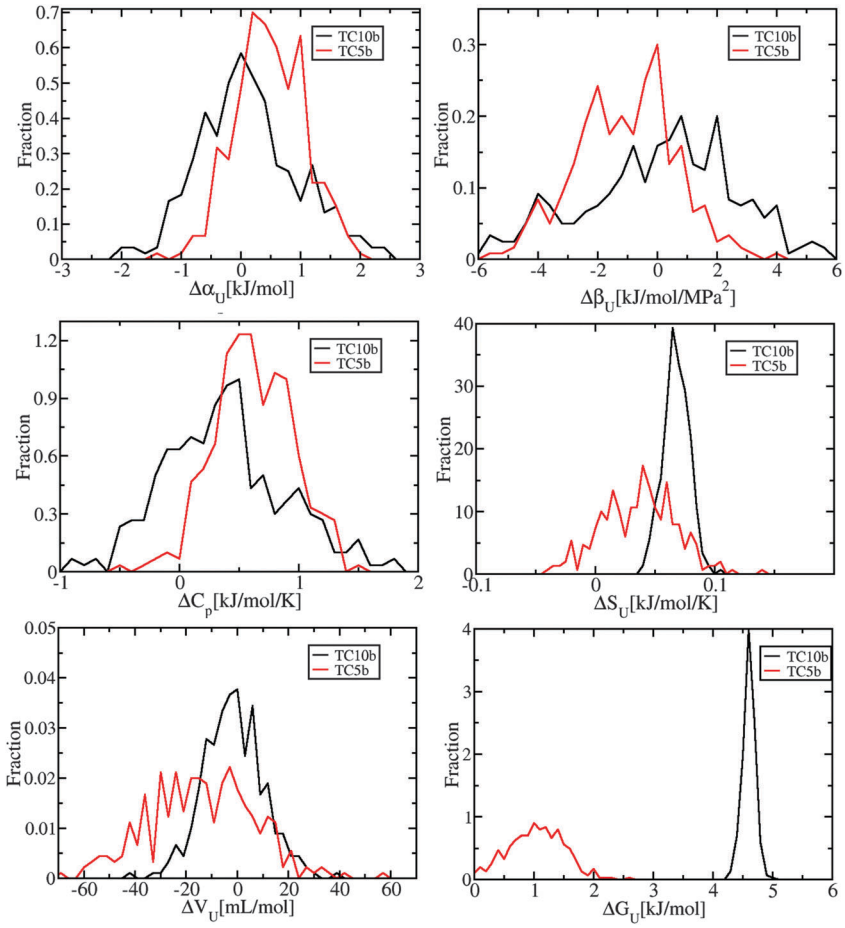

Fig. 6 Distributions of the 300 parameters obtained from bootstrapping the thermodynamic parameters for both TC10b (black) and TC5b (red).

of folding were compared. It was found that the mutations to the TC5b variant of the Trp-cage generates a relatively more ordered structure than TC5b, in addition to a more stable alpha helix and overall structure. Our result show that detailed simulations of proteins and their mutants, can reveal variations in the energy landscape and thermodynamics of proteins.

5. Appendix: simulation details

5.1 System generation and initial equilibration

All simulations were performed using the Gromacs-4.5.1 software package.[46] The AMBER99sb forcefield[12] and TIP3P water model[48] were used. An initial protein configuration was generated as linear sequence (ϕ and ψ were 180 degrees) in PyMOL.[52] Potential steric clashes in the initial configuration were relieved with an energy minimization using the steepest descent method for 500 steps. A randomized structure was generated from the linear configuration by running a 600 K MD simulation in vacuum for 500 ps. A rectangular simulation box was chosen with the closest box edge being set 1.2 nm from the protein. The box was solvated with 2863 TIP3P water molecules. As TC10b is charge neutral, no ions were added to the simulation box. Two chloride ions and one sodium ion replaced three of the water molecules to neutralize the system in Day *et al.*,[47] as the TC5b sequence is not charge neutral. The solvent was relaxed through 500 step steepest descent energy minimization and position restrained MD for 10 ps. The system was equilibrated to 300 K and 1 bar using a 10 ps MD simulation with the Berendsen thermostat and a 10 ns simulation with the Berendsen barostat. The final configuration generated in this pressure equilibration run was used as the starting structure for simulations to determine the required temperatures for a constant exchange rate, as described in the next section.

5.2 Selection of temperatures for replica exchange

In order to compute the necessary temperatures to ensure a constant exchange rate, the method outlined in García *et al.*[49] was used. The method requires a fitting of the average energy and its standard deviation for the system as a function of temperature to be calculated. For the work presented in this paper, a set of thirteen, 10 ns MD simulations were used at the temperatures 275 K, 300 K, 325 K,..., 575 K, 600 K. From the results of these simulations, a fitting of the average energy *versus* temperature $\bar{E}(T)$ and the average error *versus* temperature, $\sigma(T)$ was obtained. The energy is fitted to a cubic polynomial in T, and the standard deviation is fitted to a quadratic polynomial in T. The diagnostic simulations were performed using PME electrostatics,[50]

at constant volume with the Nose–Hoover thermostat.[71] The average energy and its standard deviation are calculated from the last 5 ns of these calculations, after the averages have reach a steady state. Given the estimates of $\bar{E}(T)$ and $\sigma(T)$ over a broad range of temperatures, an optimal set of temperatures, T_i, can be solved for iteratively that will have a given acceptance exchange rate, R_{acc}, given by the equation:[49]

$$R_{acc} = \frac{1}{2}\left[1+\mathrm{erf}\left[\frac{\bar{E}_2-\bar{E}_1}{\sqrt{2\sigma_1^2+2\sigma_2^2}}\right]\right]+\frac{1}{2}[\exp[\Delta\beta(\bar{E}_2-\bar{E}_1)+\left(\frac{\Delta\beta}{2}\right)^2(2\sigma_1^2+2\sigma_2^2)\Big]\Big]\mathrm{erfc}\left[\frac{\Delta\beta(2\sigma_1^2+2\sigma_2^2+\bar{E}_2-\bar{E}_1)}{\sqrt{2\sigma_1^2+2\sigma_2^2}}\right] \quad (5)$$

The iterative process is started by selecting the lowest desired temperature, T_1, and then solving for T_2, such that a desired value for R_{acc} is obtained. Given T_2, T_3 is estimated, and so on, until the highest desired temperature is reached. T_1 was selected to be 275.0 K, and the highest temperature, T_{50}, was calculated to be 559.4 K, with an even number of replicas ($N = 50$), and $R_{acc} = 25\%$ (Fig. 7). This method has the benefit of sampling a larger temperature range with fewer replicas than a purely geometric selection of temperatures would allow, and produces a uniform R_{acc} over all replicas. Eqn (5) assumes the energy distributions calculated in the diagnostic simulations is uniformly Gaussian following

$$P(E|E,\bar{E},\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(E-\bar{E})^2}{2\sigma^2}\right] \quad (6)$$

The rate of exchange is related to the area of overlap between two Gaussian distributions for the energies. The area of overlap is given as

$$A_{overlap} = \int_{-\infty}^{E_i} dE P(E|\bar{E}_2,\sigma_2) + \int_{E_i}^{\infty} dE P(E|\bar{E}_1,\sigma_1) \quad (7)$$

which is approximately

$$A_{overlap} \approx \mathrm{erfc}\left[\frac{\Delta E}{2\sqrt{2}\sigma_m}\right] \quad (8)$$

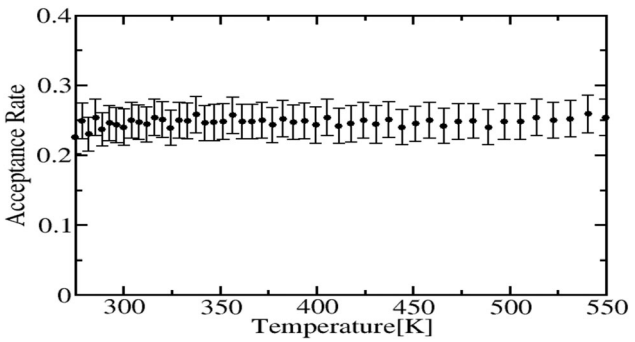

Fig. 7 Rate of acceptance with next highest temperature for replica exchange in UIC simulation with errors representing one standard deviation.

while the rate of exchange is given by

$$R_{acc} = \int_{-\infty}^{\infty} dx P(x|\bar{E}_1,\sigma_1)\left[\int_{-\infty}^{x} dy P(y|\bar{E}_2,\sigma_2) + \int_{x}^{\infty} dy P(y|\bar{E}_2,\sigma_2)\exp(\Delta\beta(x-y))\right] \quad (9)$$

Variations in R_{acc} can result due to the correlation time of the energy at various temperatures. The temperatures used for the REMD calculations of the TC10b protein are: 275.0, 278.5, 281.9, 285.5, 289.0, 292.7, 296.4, 300.2, 304.1, 308.0, 312.0, 316.1, 320.2, 324.4, 328.8, 333.2, 337.7, 342.2, 346.9, 351.7, 356.6, 361.5, 366.6, 371.8, 377.1, 382.6, 388.1, 393.8, 399.6, 405.6, 411.6, 417.9, 424.3, 430.8, 437.5, 444.3, 451.4, 458.6, 465.9, 473.5, 481.2, 489.1, 497.3, 505.6, 514.1, 522.7, 531.6, 540.7, 549.9, 559.4. Exchanges were attempted every 4 ps. A Mathematica script (temperature_assignment.nb) implementing this method is provided in the ESI.†

5.3 Bootstrap method and thermodynamic fitting

An estimate of the errors was found for each of the parameters in the Hawley equations using the bootstrap method. The bootstrap methods was chosen as the correlation time of the system was found to be about 100 ns. In contrast to other methods of error calculation, such as block averaging or taking 200 ns time blocks as was done for the helicity and folding temperature calculations, which would require the subdivision of the data set into blocks with time lengths shorter than the correlation time to generate a good estimate of the errors. Bootstrapping was desirable as it can generate a distribution of the likelihood of each parameter while ensuring adequate sampling from the simulation data. The equations used to calculate the bootstraps are described in the main paper, however, some elucidation is given as to how the equations were used here. The Maquardt method,[72] as implemented in Mathematica,[73] was used to simultaneously fit $\Delta G(P_i, T_i)$, for each replica state $\langle T, P\rangle_{T_i,V}$, described in eqn (3), to the Hawley function, described in eqn (10), *via* the thermodynamic parameters: $\Delta\alpha_U(T_o, P_o)$ (the coefficient of linear expansion), $\Delta\beta_U(T_o, P_o)$ (the isothermal compressibility), $\Delta C_p(T_o, P_o)$ (the change in heat capacity at constant pressure), $\Delta S_U(T_o, P_o)$ (the change in entropy), $\Delta V_U(T_o, P_o)$ (the change in volume), and $\Delta G_U(T_o, P_o)$ (the change in free energy). A Mathematica script was utilized to perform the calculations (given in the ESI† as bootstrap_thermodynamics.nb). Individual frames were chosen randomly from the simulation and the associated rmsd (to determine folding status of the frame *i.e.* C_α RMSD < 0.22 nm for folded or RMSD > 0.50 nm for unfolded), energy, pressure, temperature were read. For each temperature in the simulation, 600 000 configurations were computed. For each configuration, the RMSD, energy, pressure, and temperature were used. The 32 lowest temperatures were used for the fitting. In all, 60 000 data points per temperature were chosen at random from this list to comprise the underlying data set for the bootstrap.

The assignment into folded or unfolded ensembles are determined from the RMSD, with any structure with an RMSD

of less than 0.22 nm considered folded and otherwise unfolded. A cutoff of 0.50 nm was used to denote unfolded structures, eliminating intermediate structures from the calculation. The corresponding energy values were also categorized as belonging to a folded or unfolded states based on the associated RMSD value. The temperatures and pressures were averaged and the unfolded average energy was subtracted from the folded average energy to find the energy difference. The resulting values of ΔG, $\langle T_i \rangle$, and $\langle P_i \rangle$ values for each one of the fifty replicas was fitted to eqn (1). This calculation was repeated for 300 independent data sets generated at random form the whole equilibrium ensemble to estimate the errors in the fitting.

The thermodynamic relationships are defined as follows:

$$\begin{aligned}\Delta G(T,P) &= \Delta G_{\mathrm{U}} - \Delta S_{\mathrm{U}}(T - T_{\mathrm{o}}) \\ &\quad - \Delta C_{\mathrm{p}}\left[T\left(\log\left(\frac{T}{T_{\mathrm{o}}}\right) - 1\right) + T_{\mathrm{o}}\right] \\ &\quad + \Delta V_{\mathrm{U}}(P - P_{\mathrm{o}}) + \Delta\beta_{\mathrm{U}}(P - P_{\mathrm{o}})^2/10000 \\ &\quad + \Delta\alpha_{\mathrm{U}}(P - P_{\mathrm{o}})(T - T_{\mathrm{o}})/1000\end{aligned} \tag{10}$$

$$V(P,T) = \left(\frac{\partial G(T,P)}{\partial P}\right)_T \tag{11}$$

$$S(P,T) = \left(-\frac{\partial G(T,P)}{\partial T}\right)_P \tag{12}$$

$$H(T, P) = G(T, P) + TS(T, P) \tag{13}$$

$$E(T, P) = H(T, P) - PV(T, P) \tag{14}$$

Notice that the values for ΔG in the above equations contains coefficients that correct for the units used in Gromacs (kJ, bars; nm^3, *etc.*) and the units reported in Table 1 (kJ, MPa, ml mol^{-1}). The last four equations allow for various other thermodynamic quantities to be back calculated as a function of temperature. The chi-squared function that was minimized was:

$$\begin{aligned}\chi^2 &= \sum_i \left[\frac{\Delta G_i - \Delta G(T_i, P_i)}{\sigma_{\Delta G_i}}\right]^2 + \sum_i \left[\frac{x_i^{\mathrm{folded}} - x(T_i, P_i)}{\sigma_{x_i}}\right]^2 \\ &\quad + \sum_i \left[\frac{\Delta E_i - \Delta E(T_i, P_i)}{\sigma_{\Delta E_i}}\right]^2\end{aligned} \tag{15}$$

where $\sigma_{\Delta Y_i}$ is the uncertainty in the average of Y, where Y is the relevant corresponding quantity in the equation (above) for ΔG_{U}, as defined by the above equation, x the fraction folded, i the corresponding replica number, and ΔG and ΔE the changes in free energy and energy, respectively. The chi-squared function was minimized with the Newton–Raphson algorithm. The Mathematica script then outputs the found parameters to a file which may be used to construct histograms which are shown in Fig. 5 and 6.

Acknowledgements

This work was funded by the National Science Foundation through grant MCB-1050906.

References

1 L. Hou, I. Kang, R. Marchant and M. Zagorski, *J. Biol. Chem.*, 2002, **277**, 40173–40176.
2 K. Kamino, H. Orr, H. Payami, E. Wijsman, E. Alonso, S. Pulst, L. Anderson, S. O'dahl, E. Nemens, J. White, S. Sadovnick, M. Ball, J. Kayue, A. Warren, M. Mcinnis, S. Antonarakis, J. Korenberg, V. Sharma, W. Kukull, E. Larson, L. Heston, G. Martin, T. Bird and G. Schellenberg, *Am. J. Hum. Genet.*, 1992, **51**, 998–1014.
3 C. Nilsberth, A. Westlind-Danielsson, C. Eckman, M. Condron, K. Axelman, C. Forsell, C. Stenh, J. Luthman, S. Teplow, S. Younkin, J. Naslund and L. Lannfelt, *Nat. Neurosci.*, 2001, **4**, 887–893.
4 M. Bornebrook, J. Haan and R. Roos, *Amyloid*, 1999, **6**, 215–224.
5 C. Chothia and A. Lesk, *EMBO J.*, 1986, **5**(4), 823–826.
6 L. Qui, S. Pabit, A. Roitberg and S. Hagen, *J. Am. Chem. Soc.*, 2002, **124**, 12952–12953.
7 J. W. Neidigh, R. Fesinmeyer and N. Andersen, *Nat. Struct. Biol.*, 2002, **9**(6), 425–430.
8 C. Simmerling, B. Strockbine and A. Roitberg, *J. Am. Chem. Soc.*, 2002, **124**, 11258–11259.
9 J. Pitera and W. Swope, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 7582–7592.
10 J. Juraszek and P. Bolhuis, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 15859–15864.
11 L. Zhan, J. Chen and W. Liu, *Proteins*, 2007, **66**, 436–443.
12 V. Hornak, R. Abel and A. Okur, *Proteins*, 2006, **65**, 712–725.
13 M. Seibert, A. Patriksson, B. Hess and D. van der Spoel, *J. Mol. Biol.*, 2005, **354**, 173–183.
14 D. Paschek, H. Nymeyer and A. E. García, *J. Struct. Biol.*, 2007, **157**, 524–533.
15 A. Patriksson, C. Adams, F. Kjeldsen, R. Zubarev and D. van der Spoel, *J. Phys. Chem. B*, 2007, **111**, 13147–13150.
16 C. Snow, B. Zagrovic and V. Pande, *J. Am. Chem. Soc.*, 2002, **124**, 14548–14549.
17 S. Chowdhury, M. Lee, G. Xiong and Y. Duan, *J. Mol. Biol.*, 2003, **327**, 711–717.
18 Y. Chebaro, X. Dong, L. Rozita, P. Derraumaux and N. Mousseau, *J. Phys. Chem. B*, 2009, **113**(1), 267–274.
19 X. H. Huang, M. Hagen, B. Kim, R. Friesner, R. Zhou and B. Berne, *J. Phys. Chem. B*, 2007, **111**, 5405–5410.
20 R. Zhou, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 13280–13285.
21 J. Juraszek and P. Bolhuis, *Biophys. J.*, 2008, **95**, 4246–4257.
22 D. Paschek, S. Hempel and A. E. García, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 17754–17759.
23 S. Kannan and M. Zacharias, *Proteins*, 2009, **76**, 448–460.
24 J. Maupetit, P. Derraumaux and P. Tufféry, *J. Comput. Chem.*, 2010, **31**(4), 726–738.
25 D. Naduthambi and N. Zondlo, *J. Am. Chem. Soc.*, 2006, **128**, 12430–12431.
26 P. Hudaky, P. Straner, V. Farkas, G. Varadi, G. Toth and A. Perczel, *Biochemistry*, 2008, **47**, 1007–1016.
27 M. Bunagan, X. Yang, J. Saven and F. Gai, *J. Phys. Chem. B*, 2006, **111**, 3759–3763.

28 B. Barua, J. Lin, V. Williams, P. Kummler, J. Neidigh and N. Andersen, *Protein Eng., Des. Sel.*, 2008, **21**(3), 171–185.
29 W. Streicher and G. Makhatadze, *Biochemistry*, 2007, **46**, 2876–2880.
30 G. Nikiforovich, N. Andersen, R. Fesinmeyer and C. Frieden, *Proteins*, 2003, **52**, 292–302.
31 H. Neuweiler, S. Doose and M. Sauer, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 16650–16655.
32 Z. Ahmed, I. A. Best, A. V. Milchenin and S. A. Asher, *J. Am. Chem. Soc.*, 2005, **127**, 10943–10950.
33 P. Rogne, P. Ozdowy, C. Richter, K. Saxena, H. Schwalbe and L. Kuhn, *PLoS One*, 2012, **7**, 1–13.
34 J. Tian and A. García, *J. Am. Chem. Soc.*, 2011, **133**, 15157–15164.
35 P. Hudaky, P. Straner, V. Farkas, G. Varadi, G. Toth and A. Perczel, *Biochemistry*, 2008, **47**, 1007–1016.
36 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. Kleperis, R. Dror and D. Shaw, *Proteins*, 2010, **78**, 1950–1958.
37 J. Lin, B. Barua and N. Andersen, *J. Am. Chem. Soc.*, 2004, **126**, 13679–13684.
38 E. Harper and G. Rose, *Biochemistry*, 1993, **32**(30), 7605–7609.
39 D. R. Canchi, D. Paschek and A. E. García, *J. Am. Chem. Soc.*, 2010, **132**, 2338–2344.
40 D. R. Canchi and A. E. García, *Biophys. J.*, 2011, **100**, 1526–1533.
41 D. R. Canchi and A. E. García, *Annu. Rev. Phys. Chem.*, 2013, **64**, 273–293.
42 C. Jimenez-Cruz, G. I. Makhatadze and A. E. García, *Phys. Chem. Chem. Phys.*, 2011, **13**, 17056–17063.
43 Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**(1–2), 141–151.
44 S. Hawley, *Biochemistry*, 1971, **10**(13), 2436–2442.
45 B. Efron, *Ann. Stat.*, 1977, **7**(1), 1–26.
46 B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.
47 R. Day, D. Paschek and A. E. García, *Proteins*, 2010, **78**, 1889–1899.
48 W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey and M. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
49 A. E. García, H. Herce and D. Paschek, *Annu. Rep. Comput. Chem.*, 2006, **2**, 83–95.
50 T. Darden, Y. Darrin and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
51 B. Cooke and S. Schmidler, *J. Chem. Phys.*, 2008, **129**, 164112–164129.
52 L. Schrodinger, *The PyMOL Molecular Graphics System, Version 1.1*, Schrodinger, LLC., 2008.
53 H. Berendsen, J. Postma, W. van Gunsteren, A. DiNola and J. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3691.
54 L. Smeller, *Biochim. Biophys. Acta*, 2002, **1595**, 11–29.
55 D. Frenkel and B. Smit, *Understanding Molecular Simulation*, Elsevier, San Diego, CA, 2nd edn, 2002.
56 A. E. García and K. Sanbonmatsu, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 2782–2787.
57 X. Daura, R. Suter and W. van Gunsteren, *J. Chem. Phys.*, 1999, **110**, 3049–3055.
58 G. Hummer, S. Garde, A. E. García, M. Paulaitis and L. Pratt, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 1552–1555.
59 J. Roche, J. Caro, D. Norberto, P. Barthe, C. Roumestand, J. Schlessman, A. E. García, B. Garcia-Moreno and C. Royer, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**(18), 6945–6950.
60 T. Ghosh, A. E. García and S. Garde, *J. Am. Chem. Soc.*, 2001, **123**, 10997–11003.
61 T. Ghosh, A. E. García and S. Garde, *J. Chem. Phys.*, 2002, **116**, 2480–2486.
62 T. Ghosh, A. E. García and S. Garde, *J. Phys. Chem. B*, 2003, **107**, 612–617.
63 K. Mok, L. Kuhn, M. Goez, I. Day, J. Lin, N. Andersen and P. Hore, *Nature*, 2007, **447**, 106–109.
64 S. Showalter and R. Bruschweiler, *J. Chem. Theory Comput.*, 2007, **3**, 961–975.
65 R. Best and G. Hummer, *J. Phys. Chem. B*, 2009, **113**, 9004–9015.
66 J. Brandts, R. Oliveira and C. Westort, *Biochemistry*, 1970, **9**, 1038–1047.
67 G. Panick, G. Vidugiris, R. Malessa, G. Rapp, R. Winter and C. Royer, *Biochemistry*, 1999, **38**, 4157–4164.
68 H. Herberhold and R. Winter, *Biochemistry*, 2002, **41**, 2396–2401.
69 H. Herberhold and R. Winter, *Biochemistry*, 2002, **41**(7), 2396–2401.
70 D. Paschek and A. E. García, *Phys. Rev. Lett.*, 2004, **93**, 238106.
71 S. Nosé, *J. Chem. Phys.*, 1984, **81**, 511–520.
72 D. Marquardt, *J. Soc. Ind. Appl. Math.*, 1963, **11**(2), 431–441.
73 Wolfram, *Wolfram Mathematica 7*, 2008.
74 L. Wafer, W. Streicher and G. I. Makhatadze, *Proteins*, 2010, **78**(6), 1376–1381.