PCCP

Cite this: *Phys. Chem. Chem. Phys.*, 2011, **13**, 15166–15174

www.rsc.org/pccp

PAPER

The protein folding transition-state ensemble from a Gō-like model†

Athi N. Naganathan[a] and Modesto Orozco*[abc]

Received 29th March 2011, Accepted 20th June 2011
DOI: 10.1039/c1cp20964g

Characterizing the structure of transition states (TS) is a first step towards understanding two-state protein folding mechanisms. However, a direct experimental characterization of these states is challenging and indirect information derived from protein engineering methodologies (ϕ-value analysis) is often difficult to interpret. We present here a theoretical study on the nature of the transition state ensemble for three representative proteins covering the major structural classes using a mean-field C_α-based Gō-model. We identify that transition state ensembles are dominated by local contacts, indicating that most non-local contacts form only upon crossing the macroscopic folding free energy barrier. We demonstrate that the mean ϕ-value corresponds to the fraction of stabilization energy gained at the barrier-top in two-state-like systems, and that it depends monotonically on the stability conditions. Furthermore, we show that there is a fundamental connection between small destabilization and large ϕ-values that in turn depends on the location of the mutated residue in the structure. These results that are in agreement with the recent empirical findings highlight the importance of local energetics in determining folding mechanisms.

Introduction

Understanding the intricate mechanistic details of protein folding is one of the grand challenges in biological chemistry. Popular model systems in this regard are two-state-like proteins, *i.e.* proteins in which only two apparent macrostates, folded and unfolded, are populated under all conditions separated by a large free energy barrier. The stability of the folded states is of the order of just a few *RT* arising out of cancellation between large energetic and entropic terms.[1,2] Folded states are well defined from both structural and dynamical points of view,[3] but much less is known about the features of the unfolded state. The least characterized and possibly the most relevant to the folding mechanism of two-state proteins are the transition states (TS), *i.e.* barrier top structures, which are by definition poorly populated.

The experimental methodology to access information on the structure of the TS for two-state proteins was developed primarily by Fersht and co-workers and is termed the ϕ-value analysis.[4,5] In this methodology, single point mutations that typically correspond to side-chain truncations are performed on the wild-type protein. The equilibrium constants (K_{eq}) and the rate constants (k) are measured for both the wild-type (wt) and the mutants (mut). The ϕ-value is calculated as:

$$\phi = \frac{\Delta\Delta G_{\text{fol}}^{\text{mut-wt}}}{\Delta\Delta G_{\text{eq}}^{\text{mut-wt}}} = \frac{\ln(k_{\text{fol}}^{\text{wt}}/k_{\text{fol}}^{\text{mut}})}{\ln(K_{\text{eq}}^{\text{wt}}/K_{\text{eq}}^{\text{mut}})} \qquad (1)$$

where fol stands for folding and eq for the equilibrium, and in essence is a two-point Brønsted analysis. ϕ-Values of 0 and 1 would therefore correspond to scenarios in which the folding TS are as unstructured as the unfolded state or as folded as the native state, respectively. Accordingly, residues with high ϕ-values are seen as 'folding nucleus' that acquires a native-like structure early and around which the protein folds.[5] In practice, however, the experimental ϕ-values are often fractional numbers making their interpretation challenging.[6] They are further influenced by additional factors such as experimental noise,[7] unfolded state structure,[8] and even by laboratory-specific experimental procedures.[9] Given these issues, questions on the physical meaning of fractional ϕ-values arise.

There are, in principle, two possible interpretations for fractional ϕ-values. The first assumes a single TS structure with a partial degree of structure formation in the mutated site.[4] The second scenario is one in which the TS is not a single structure but an ensemble (the transition state ensemble, TSE), with various degrees of structure formed at the mutated site whose average corresponds to the observed experimental number.[10–12] Distinguishing between both possibilities is not

[a] *BSC-IRB Joint Research Program in Computational Biology, Barcelona Supercomputing Center, Torre Girona, C/Jordi Girona 31, Barcelona 08034, Spain. E-mail: anarayan@bsc.es*
[b] *IRB-BSC Joint Research Program in Computational Biology, Institute for Research in Biomedicine, Parc Cientific de Barcelona, C/Baldiri Reixac 10, Barcelona 08028, Spain. E-mail: modesto.orozco@irbbarcelona.org*
[c] *Departamento de Bioquímica, Facultat de Biologia, Avgda Diagonal 647, Barcelona 08028, Spain*

† Electronic supplementary information (ESI) available: Additional figures supporting the results. See DOI: 10.1039/c1cp20964g

a trivial issue to tackle experimentally as the solution lies in being able to probe the least-populated species along a reaction mechanism (the TS) at the single-molecule level. In a recent study, Naganathan and Muñoz performed a global empirical analysis of the effects of mutations on the folding rate and stability on more than 800 mutations from 24 different proteins covering all three structural classes and that span the microsecond–millisecond folding regime.[13] Their results presented strong evidence for the following: (a) folding transition states have little tertiary structure and are dominated by local interactions—suggestive of a "local-first" mechanism in protein folding, (b) transition states move monotonically with protein stability in accordance with Hammond behaviour and hence reference conditions are crucial in comparing ϕ-values, (c) under a suitable reference condition the ϕ-values cluster around 0.36, and finally (d) exposed residues in a folded protein result in smaller destabilization and have higher ϕ-values compared to buried ones, contrary to the 'folding nucleus' hypothesis. Overall, it was concluded that transition states are better seen as ensembles in tune with the energy landscape theory of protein folding.[14]

However, despite the richness of information derived from the massive analysis of ϕ-values we cannot ignore the fact that this methodology, as with any other experimental approach to the study of two-state folders, is limited by the impossibility to track at the atomic level the folding of a protein. Theoretical methods then provide a viable recourse to complement experimental techniques. In this regard, all-atom molecular dynamics (MD) simulations have proven to be quite competent in capturing native-state dynamics,[3,15] to test folding mechanisms,[16,17] identify multiple folding routes[18] and to make experimentally testable predictions.[19] Unfortunately, despite recent technical advances, MD is still limited to the study of folding processes on the micro-second time scale. Moreover, it is computationally prohibitive to simulate multiple folding–unfolding events, making it necessary the use of simpler Hamiltonians and coarse-grained representations of proteins. In particular, the mean-field C_α Gō-models[20] invoking the minimal frustration principle[14] represent a simple, but robust alternative. Apart from capturing the basic physics behind folding, they have been successfully used to predict folding mechanisms,[21,22] protein assembly,[23] and even allosteric transitions.[24]

Here, we employ a mean-field C_α-based Gō-model to gain information on the nature of the transition state ensembles and on the meaning of fractional ϕ-values. To this end, we characterize the folding behaviour of three proteins representative of the three main structural classes (α, α/β and β), finding strong evidence for the dominance of local-interactions in the TSE, the Hammond behaviour, diversity of interactions in the TSE and the intricate connection between folding energetics and position in the structure of the mutated residue.

Methods

Parameterization of the Gō-model

The C_α-based Gō-model considered here was parameterized along the same lines as proposed by Onuchic and co-workers[25] and quite successfully used in a variety of studies.[26–29] The potential energy function (Hamiltonian) used is as shown below:

$$V = V_{\text{bonded}} + V_{\text{angle}} + V_{\text{dihedral}} + V^{\text{native}}_{\text{non-bonded}} + V^{\text{non-native}}_{\text{non-bonded}} \tag{2}$$

$$= \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} \{K_\phi^{(1)}[1-\cos(\phi-\phi_0)] + K_\phi^{(3)}[1-\cos 3(\phi-\phi_0)]\} + \sum_{\text{native}} \varepsilon\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right] + \sum_{\text{non-native}} \varepsilon\left(\frac{\sigma_0}{r_{ij}}\right)^{12} \tag{3}$$

where r, θ and ϕ are the bond length, angle and dihedral angle, respectively, in the first three terms. The corresponding subscripts '0' stand for values in the PDB structure. The fourth term corresponds to the Lennard-Jones-like (LJ) stabilization energy represented as a 12-10 function that acts only on those contacts present in the native-state. Here, the r_{ij} and σ_{ij} identify the distance between two atoms i and j in any snapshot and in the native-state, respectively. The fifth term is an excluded volume function that energetically disfavours any close non-native contact (σ_0 = 4 Å). In its current implementation, native-contacts were identified with a 5 Å heavy-atom cut-off excluding up to $i - i + 3$ sequential neighbours. Nearest-neighbour energy terms are defined by: $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\phi^{(1)} = \varepsilon$, and $K_\phi^{(3)} = 0.5\varepsilon$ where ε sets the energy scale and is independent of the residue identity (see ref. 25 and 30 for additional details). A Langevin dynamics approach was used to simulate folding–unfolding[31] as implemented in GROMACS 4.0.2[32] with a friction constant of 1 ps^{-1}. For each protein and at each temperature 10^8 steps were run with an integration time-step of 0.5 fs. Longer simulations (10 × 10^8 steps) were carried out at temperatures of interest to generate sufficient statistics.

ϕ-Values

ϕ-Values were calculated as the fraction of contacts in the TSE:

$$\phi_i^Q = \frac{\langle n_i \rangle_{\text{TSE}}}{N_{i,F}} \tag{4}$$

where the numerator is the mean number of native contacts for residue i in the TSE with $N_{i,F}$ being the corresponding number in the native-state. Free-energy perturbation calculations were also performed assuming that mutations uniformly affect the interacting residues with little effects on the folding mechanism. This was implemented by reducing the LJ energies of the interacting partners of a particular residue by 50% and the resulting ϕ-values were calculated as:[10]

$$\phi_i^P = \frac{\ln\langle e^{\Delta E_i/k_B T}\rangle_{\text{TSE}}}{\ln\langle e^{\Delta E_i/k_B T}\rangle_F} \tag{5}$$

where ΔE is the difference in the potential energy relative to the wild-type, k_B is Boltzmann's constant, and T is the simulation temperature. The brackets signal averaging over time. The magnitude of the calculated ϕ-values does not depend on the percentage of interaction energy reduced (see Fig. S1, ESI†). There is also a good agreement between ϕ-values calculated from eqn (4) or (5) and when unfolded state information is included (ϕ^Q is therefore referred to as ϕ unless otherwise mentioned). However, using a fixed reference such as the PDB structure aids in identifying trends as we are interested in just the degree of structure formed at the transition state. The local/total neighbour ratios were calculated as in the original work[13] from the PDB structure of the proteins by using a 6.8 Å cut-off and with a binary description of the contact-map: if the interacting residues are less than four residues apart in sequence, they are considered as local and otherwise non-local (total = local + non-local).

Bayesian analysis of the transition path

One could think of several reaction coordinates (RC) including the fraction of native contacts (Q),[33,34] root mean square deviation to the native structure (RMSD), fraction of native-dihedrals *etc.* to construct free-energy surfaces. Of late, p_{fold} or committor probability[35] is being increasingly used to test the robustness of the constructed free-energy surfaces in identifying transition state ensembles. Free-energy surfaces on chosen reaction coordinates were built from histogram analysis methods using unbiased trajectories where many folding/unfolding events were recorded.

Here, we use Hummer's Bayesian analysis[36–38] to check for the suitability of using a particular reaction coordinate to identify the TSE. This methodology is based on the inherent differences in distribution between the equilibrium ensemble and the transition path (TP) ensemble—defined as the series of steps consecutive in time that connect the folded ensemble to the unfolded ensemble without returning and *vice versa*—that are by definition rarely populated in equilibrium. The probability of being in a transition path given a particular reaction coordinate value (r) can be expressed as

$$p(\text{TP}|r) = \frac{p(r|\text{TP})p(\text{TP})}{p_{\text{eq}}(r)} \qquad (6)$$

Assuming a suitable reaction coordinate, each of the variables on the right-hand side of the above expression can be easily extracted from long equilibrium simulations involving multiple transitions between folded and unfolded states. Here, $p(r|\text{TP})$ is the probability of a particular reaction-coordinate value to be in a transition path which is obtained by identifying all the possible TPs and calculating the probability of r to be in the corresponding region. $p(\text{TP})$ is a constant defining the fraction of time the protein resides in the TP (*i.e.* the number of frames defined as falling within the TP over the total number of frames) while $p_{\text{eq}}(r)$ is the equilibrium distribution along the coordinate r. As demonstrated by Hummer,[36] a good reaction coordinate results in a single peaked $p(\text{TP}|r)$ distribution with a theoretical maximum of 0.5 (*i.e.* $p_{\text{fold}} = 0.5$). Both of the above requirements are fulfilled for Q and RMSD and we therefore chose these coordinates to identify TSE. In other words, we performed two independent Bayesian analyses using Q and RMSD as the reaction coordinate and selected structures that resulted in a high $p(\text{TP}|r)$ for both as putative TSE members. The peak of the $p(\text{TP}|r)$ always agreed well with the free-energy barrier top for both the coordinates (see Fig. S2, ESI†, for example). This enabled us to set a highly stringent threshold of ~±0.02 in Q-units and ±1–2 Å in RMSD on either side of the barrier top to identify TSE. The cut-offs in Q and RMSD used to identify folded/unfolded-ensembles for the transition path analysis are: 0.85/0.27 (in Q-units) and 2.2/15.2 Å for Im9, 0.86/0.19 and 3/16 Å for Ubq, and 0.8/0.18 and 4/20 Å for SH3. The projection on Q and RMSD alone does not guarantee that additional coordinates are not needed. However, with a two dimensional density map and a strict Bayesian analysis we believe that the generated TSE is well-defined.

Proteins studied

To explore the effect of topology on folding mechanisms, we studied one protein from each of the three major structural classes using the mean-field Gō-model. They are: Im9 (all-α, 1IMQ), Ubiquitin (α–β, 1UBQ), and PI3K SH3 (all β, 1PNJ) (Fig. 1). The three proteins have similar protein lengths (86, 76, and 83 residues, respectively) and therefore similar number of contacts in the native state (199,190, and 191, respectively, with a 5 Å cut-off). All of them have been studied extensively by experimental techniques and are known to fold in a two-state-like mechanism,[39–41] something that is well reproduced in the simulations. This can be seen, for example, in the sharp transition between folded and unfolded ensembles at the midpoint temperature in Fig. 2a and from the singly-peaked heat capacity thermogram in Fig. 2b. The distribution of native-contacts was also generated for the one-state protein BBL (2CYU, 39 residues in length, 48 contacts using a 5 Å cut-off) employing the same simulation methods as above.

Results

Evidence for a 'local-first' folding mechanism

Fig. 3a–c plots the free-energy surface as a function of Q at the respective midpoint temperatures identified by the peak of the thermogram (Fig. 2b). Two wells are clearly identified corresponding to the folded (high Q) and unfolded (low Q) structures, with barrier tops located at Q-values of 0.52 (Im9), 0.48 (Ubq), and 0.47 (SH3), respectively. They agree very well with the location of the single peaks of $p(\text{TP}|Q)$ along Q (0.51 (Im9), 0.47 (Ubq) and 0.46 (SH3), see insets in panels Fig. 3a–c), *i.e.* the structures with the highest free-energy as identified by Q correspond to the peak of the $p(\text{TP}|Q)$ distribution. We note that Q alone was insufficient to capture the true TSE as it resulted in structures of high RMSD that did not present a p_{fold} or $p(\text{TP}|\text{RMSD})$ of ~0.5. This necessitated

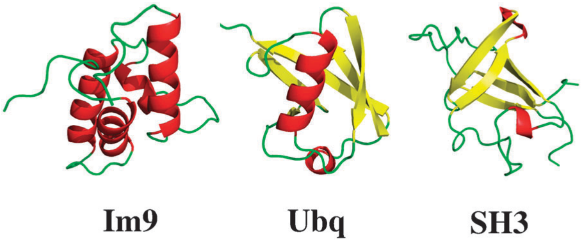

Fig. 1 Proteins studied.

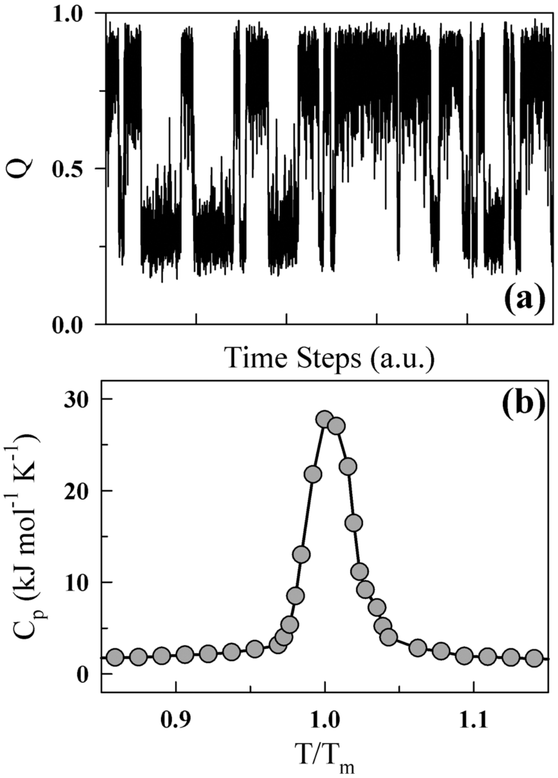

Fig. 2 Two-state-like behaviour. (a) Fraction of native contacts (Q) of Im9 as a function of simulation time showing sharp transitions between folded and unfolded wells corresponding to high and low Q values, respectively. (b) The single-peaked heat capacity thermogram of Im9 with T_m being the peak heat capacity temperature.

the use of a two-dimensional Bayesian analysis to identify the putative TSE members (see Methods and Fig. S2, ESI†). The α-helical Im9 has a smaller free-energy barrier compared to Ubq and SH3 in qualitative agreement with experimental relaxation rates.[39]

We find a significant correlation between the probability of finding a particular contact in the TSE and the sequence separation between interacting residues (Fig. 3d–f), *i.e.* larger the sequence separation the lower is the probability of finding a native-like contact. The correlation is stronger for the all-α protein and much smaller for the all-β protein as expected from simple topological considerations.[42,43] Some native contacts at low sequence separation are already present in the unfolded states of the respective proteins hinting at the importance of unfolded structure in determining the TSE and hence folding mechanisms. These observations suggest that for proteins with significant α-helical content local interactions dominate the TSE and that most long-range interactions are formed only upon crossing the macroscopic barrier. In other words, the TSE of all-α and α–β proteins lacks well-defined or stable tertiary interactions thus resembling a molten-globule.

The transition-state ensemble and Hammond behaviour

The analysis above refers to mean properties of the TSE, but how large is the underlying distribution and how does it depend on protein stability? These questions open up the subject of whether it is appropriate to envision the TS as a single structure with partially formed interactions[5] or as a large heterogeneous ensemble[10,12] and whether Hammond behaviour is a constituent property of proteins.

Characterization of protein folding with the fraction of native contacts aids in reducing the complexity of the problem but can obscure the structural diversity of the TSE. This is clearly seen in the distribution of potential energy of the barrier top TS structures (Fig. 4). It reveals a broad distribution whose corresponding native contacts fluctuate significantly between folded- and unfolded-like conformations

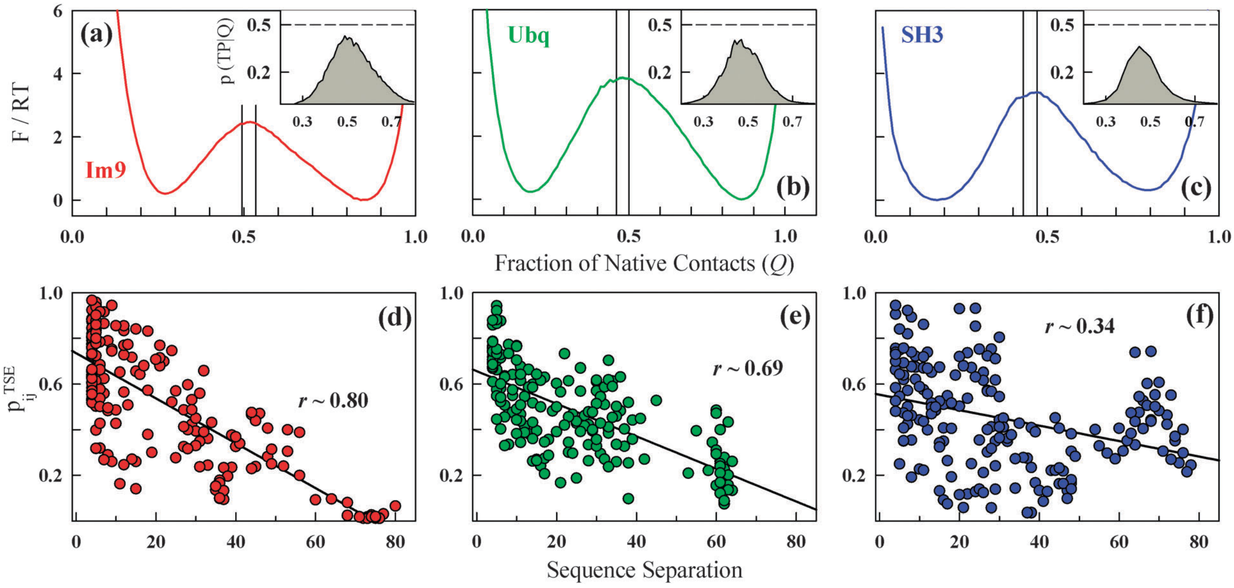

Fig. 3 Evidence for a 'local-first' folding mechanism. (a–c) Free energy profiles as a function of fraction of native contacts (Q) for the three proteins (Im9—red; Ubq—green; SH3—blue). The vertical lines signal the ∼±0.02 Q-units swath used together with the RMSD-based criteria from a Bayesian analysis for the identification of transition-states. Inset: Probability of being in a transition path given Q. (d–f) The probability of finding a contact between residues i and j in the transition state ensemble as a function of the sequence separation (lines—linear regression fit with correlation coefficient r).

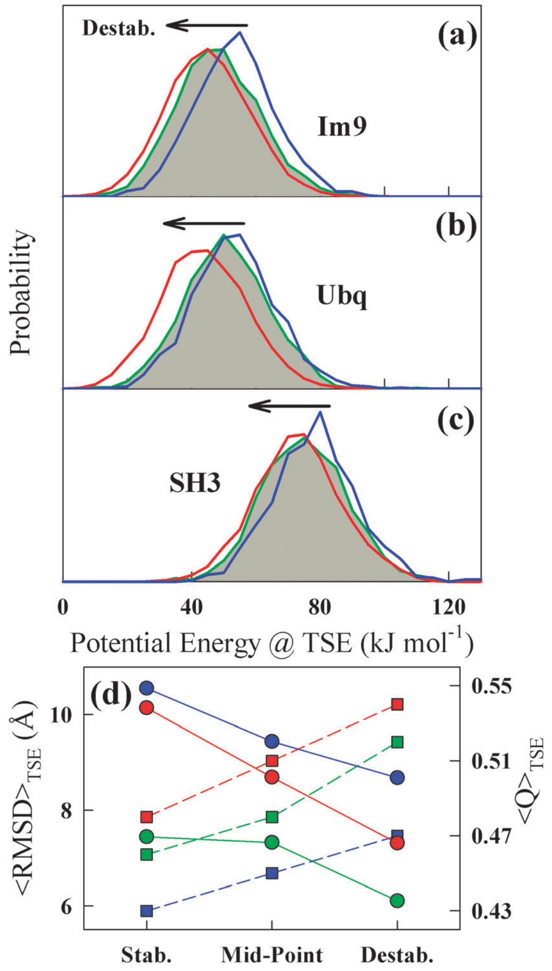

Fig. 4 The transition-state ensemble and Hammond behaviour. (a–c) Distribution of potential energy for the transition state ensemble of the three proteins under stabilizing (blue), mid-point (green with gray fill) and destabilizing conditions (red), highlighting the movement of the TSE. (d) The mean RMSD (circles and left-axis) and the mean fraction of native-contacts (squares and right-axis) for Im9 (red), Ubq (green) and SH3 (red) at different values of stability.

(see Fig. 5 at the midpoint conditions for the three proteins). The mean RMSD of the TSE is also seen to vary from 11 to 6 Å depending on the protein-type and the conditions (Fig. 4d). An ensemble with heterogeneous sets of interactions formed is therefore a much more appropriate picture of the TS than a single expanded structure. Note that this is not contradictory to the fact that well-defined ϕ-values can be obtained for individual residues. This is shown in Fig. S3 (ESI†), where irrespective of the computation protocol used (see Methods) we find well defined ϕ-values, which could be erroneously interpreted as evidence of a single TS structure with partial structure formation. Thus, we can conclude that a single ensemble experimental measure (like the ϕ-value) can in reality mask a large underlying structural distribution. One should therefore proceed with caution in structure generation methods based on, for example, standard restrained MD simulations with ϕ-values or NMR parameters, as the variance of the underlying distribution is often not known *a priori*.

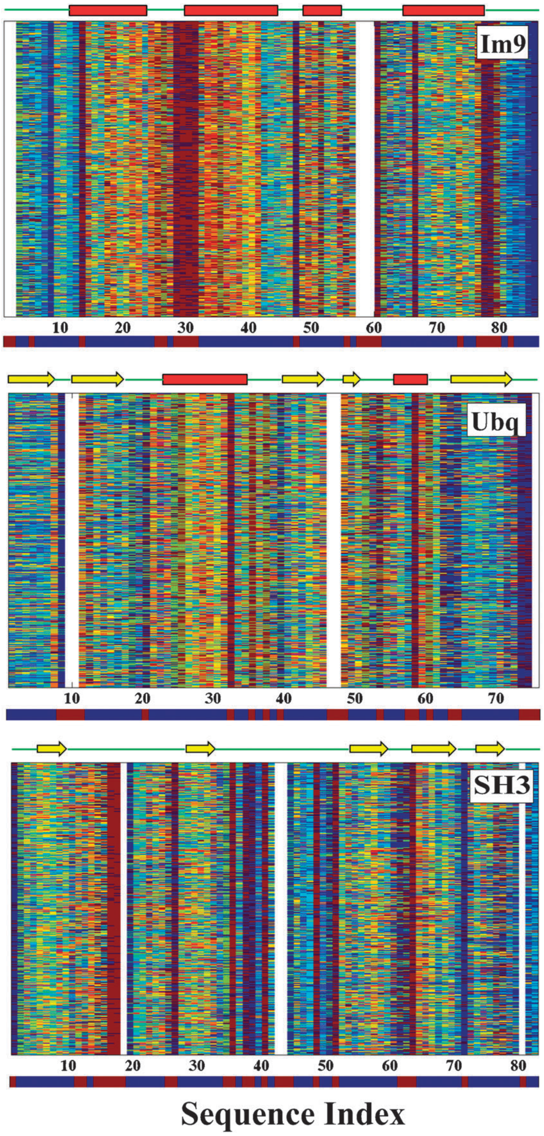

Fig. 5 Fraction of native contacts in a 500-snapshot transition-state ensemble. The colour coding for the fraction goes continuously from blue (0) through green (0.5) to red (1). The uncoloured regions correspond to residues in which no interacting partners were identified within the contact cut-off of 5 Å in the native-state (highly solvent-exposed sites) and hence the ϕ-values are undefined. The cartoon on the top of each of the TSE-map corresponds to the secondary structure as a function of the sequence (helices—red cylinders; beta strands—yellow arrows; loops—green). Note the pattern of ϕ-values in the helices of Im9. The cartoon immediately below the TSE-map highlights the residues identified by the local/total neighbour ratio greater than 0.5 in red and the rest in blue.

The mean position of the potential energy distribution is not unique, but changes with protein stability. Under highly stabilizing conditions the ensemble is more unfolded and it gets progressively folded-like (*i.e.* lower potential energy) upon decreasing the folded-state stability, in agreement with the Hammond postulate (see similar movements for ϕ-values in Fig. S4, ESI†). This behaviour is also reflected in the RMSD of the TSE with respect to the native-state (that decreases with increasing destabilization) and also the fraction of native contacts (which increase; Fig. 4d). These results are the computational equivalent of the changes in ϕ-values with temperature as seen in CI2[44] and WW-domain,[45] analysis of curved Chevron plots in U1A[46,47] and mutants of S6,[48] the multiple-perturbation analysis of tendamistat,[49,50] and with the large-scale empirical analysis of 24 different proteins with linear-limbed chevron plots.[13] Earlier analytical models of folding also provided strong evidence for this behaviour.[51] Taken together, these results suggest that the movement of the TSE with stability is a ubiquitous behaviour in small single domain proteins.

The movement of the TSE as a function stability raises another interesting question: is the folding mechanism is still conserved at different conditions? This can be inspected by calculating the average probability of contact formation in the TSE: the correlation between the probability of contact formation in the TSE and sequence separation (as in Fig. 3d–f) decreases when moving from stabilizing to destabilizing conditions (for Im9, this goes from 0.81 (stabilizing) to 0.78 (destabilizing); for Ubq from 0.74 to 0.66; and for SH3 from 0.34 to 0.29). The probability of native contacts as a function of sequence separation for the SH3 domain is almost flat and so the correlation coefficients do not have much meaning in this case. The decreasing correlation with increasing destabilization points to a higher incorporation of non-local contacts in the TSE due to the movement of the barrier top towards the native state. They suggest that the global mechanism where local-contacts form first remains conserved with changing stability, though to a lesser extent for the all-β SH3 domain. Taken together, it reveals a picture of a quite malleable TSE that responds continuously to stability changes in a Hammond-like behaviour with different ensembles populated at different conditions.

ϕ-Values and the Brønsted plot

The parameters available from experiments are the ϕ-values, or more precisely, the changes in equilibrium stability and relaxation rates upon mutation. What general features can be identified from a conventional Brønsted analysis or from the stability dependence of ϕ-values from the current data-set?

As a first step to look for global patterns in ϕ-values we have assumed that mutations affect the energetics of the interacting residues uniformly (see Methods). It reveals a direct correlation between the degree of destabilization induced by the mutations and the number of residue neighbours in the native state (Fig. 6a). The consequence is that solvent-exposed residues that have fewer neighbours will result in smaller destabilization upon mutational perturbation. The computed Brønsted plot (Fig. 6b) shows a high correlation between the destabilization energy ($\Delta\Delta G_{\mathrm{eq}}$) and the folding free energy ($\Delta\Delta G_{\mathrm{fol}}$), with an average ϕ-value of around 0.5. This is strikingly similar to the experimental Brønsted plots[13] but with a slightly larger average value. The intrinsic correlation between $\Delta\Delta G_{\mathrm{eq}}$ and $\Delta\Delta G_{\mathrm{fol}}$ is therefore a fundamental feature of protein folding transition states. In other words, with an idea of the experimental $\Delta\Delta G_{\mathrm{eq}}$ and mean ϕ-value alone one could predict the $\Delta\Delta G_{\mathrm{fol}}$ without resorting to any structural calculations using mutants. These correlations also indicate that native-states carry sufficient information on folding mechanisms supporting the use of Gō-models to study folding behaviours.

A plot of ϕ-values as a function of the destabilization energy provides a clearer picture of their dependence on stability—ϕ-values span the entire range from 0 to 1 under small destabilizations approaching a value of ~0.5 under highly destabilizing conditions (Fig. 6c and Fig. S5, ESI†). The correlation between the degree of destabilization and the number of neighbours indicates that residues that are solvent-exposed (and hence fewer neighbours or higher local/total ratio; Fig. S6. ESI†) will result in smaller destabilization. Taken together with the distribution of ϕ-values in Fig. 6c, it suggests that the intrinsically small destabilizations induced at exposed residues result in a large variability in the ϕ-values. As a result, ϕ-values close to 1 (*i.e.* fully formed) are more probable at exposed sites than buried ones as the energetic/structural environment is easily satisfied during folding by the ordering of just a few residues in contrast to buried ones. The large variability at lower destabilizations is due to the higher discretization in the contact calculation, *i.e.* the total number of contacts made by an exposed residue is small and hence the ϕ-value (defined as in eqn (4)) fluctuates significantly upon the addition/deletion of even a single extra contact in the TSE, and to a lesser extent because the structural environments of exposed sites are more

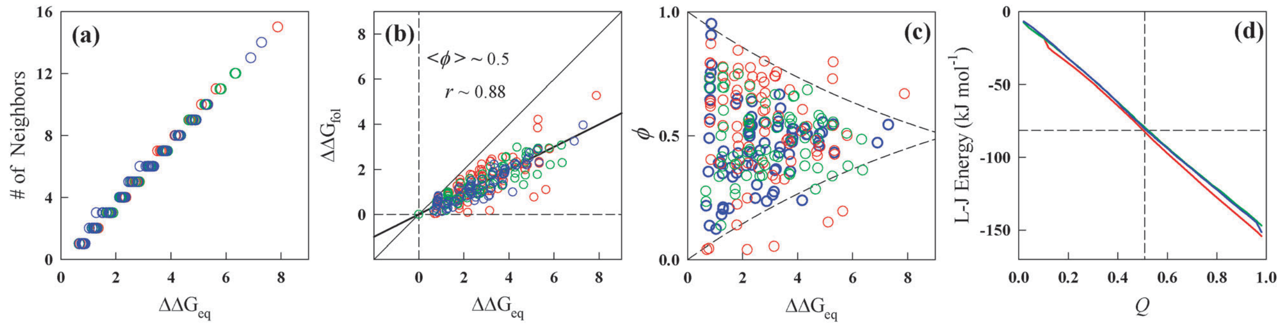

Fig. 6 ϕ-Values and the Brønsted plot. Energy units are in RT unless otherwise specified. Colour coding is maintained from Fig. 4. (a) Number of interacting neighbours as a function of the induced destabilization from the free-energy perturbation analysis. (b) Brønsted plot revealing a highly correlated distribution with a mean ϕ-value of 0.5. (c) Dependence of ϕ-values on the induced destabilization. The dashed lines are shown to guide the eye with more than 95% of the data falling within their defined limits. (d) Lennard-Jones stabilization energy as a function of the reaction co-ordinate Q. The horizontal dashed lines signal the gain in ~50% of the stabilization energy at the barrier top.

diverse compared to buried ones.[13] In experiments, however, ϕ-values for residues with lower destabilizations have significant experimental error and hence are quite unreliable, resulting in the large spread at lower destabilizations.[52] The narrowing down at large destabilizations has mainly to do with the nature of the environment of buried residues that are quite uniform in terms of the number of interacting residues.

The variation of the ϕ-values (eqn (4)) with respect to the sequence index for the individual snapshots selected from the TSE at the midpoint temperature displays a highly patterned map (labelled the TSE-map; Fig. 5 and Fig. S6, ESI†) for the three proteins. More native-like residues are located predominantly in the loop regions where there are fewer contacts (as they are more exposed). Helices are better formed than β-strands at TSE as they are dominated by local interactions. A local/total neighbour ratio of equal or greater than 1/2 (see Methods) captures the folded-like residues in the TSE, suggesting that a simple analysis of the native structure (from PDB) can predict features of transition-states to a first approximation. Finally, we also observe a characteristic periodicity in the distribution of ϕ-values for α-helical residues that is particularly more evident in Im9 (Fig. 5). This interesting finding results from the exposed/buried periodicity of residues in α-helical regions of folded proteins again conforming to experimental findings.[53]

Discussion

Our results provide an ensemble view of the nature of the transition state in two-state folding proteins that is constantly gaining support from simulations.[10–12] We show that the 'transition state' is not a privileged structure sampled during the folding process, but constitutes quite a diverse ensemble. Previous works had considered only the distribution of structures that could encompass a TSE, but here we go further to show that the distribution depends consistently on the stability conditions, as expected from Hammond's principle. The structural diversity at the transition state ensemble is also able to justify in a very simple way fractional ϕ-values (resulting from a combination of structures where a particular residue is in native environment and while in others it is unfolded-like), arguing against the traditional interpretation of these parameters in the context of a well-defined transition state structure. The finding that high ϕ-values correspond to solvent-exposed sites on a protein is strong evidence against the 'nucleation–condensation' mechanism.[5,54] This is consistent with the fact that an ensemble description provides a better agreement with experimental ϕ-values than when a single optimal 'folding nucleus' is considered.[28,55] Residues with high ϕ-values also exhibit large aggregation propensities adding an extra dimension to the folding/mis-folding problem.[56,57]

Topological effects

The relative abundance of local contacts in the TSE we find here has also been previously observed in other independent works: in the folding of 3D-lattice representation of proteins,[58] and for CI2 (α–β protein) and cytochrome *C* (all-α) from a C_α-based Gō-model.[25,59,60] The prevalence of local-contacts does not mean that non-local contacts are completely absent, but that their frequency of occurrence in the ensemble is much smaller that moreover depends on the stability conditions. Interestingly, the larger fraction of non-local contacts we obtain here for PI3K SH3 is also seen in the TSE of src-SH3 domain (all-β),[25] which suggests that all-β proteins exhibit a more complex folding mechanism possibly due to long-range correlations required for the ordering of β-sheets in proteins. This is evident in the relative contact order calculated from TSE structures (from Fig. 3d–f) resulting in values of 0.08, 0.12 and 0.14, respectively, for Im9, Ubq and SH3, implying a larger incorporation of non-local contacts upon increasing the β-content. The folding rates should therefore scale accordingly at the midpoint as the corresponding barrier heights are 2.5, 3.9 and 3.4 *RT* units, with Im9 folding being the fastest.

The more complex folding of SH3 is apparently at odds with the simulation results for all-α and α–β proteins that seem to fold in a simpler fashion by local ordering. However, as discussed below, the simulated ϕ-values is an upper limit to the degree of structure formed in the TSE and hence still consistent with a 'local-first' mechanism since the lower the degree of structure the higher the expected correlation with sequence separation. This mechanism also provides an intuitive avenue to approach the protein folding problem that has been quite successfully implemented in the Rosetta algorithm to predict 3D structures[61] and in the mean-field Ising-like models that serve as powerful tools in characterizing folding behaviour.[62] The fact that just backbone topology and local interactions are sufficient to generate most of the known stable folds has also been shown from simple homopolymer models[63] further supporting our conclusions.

Comparison with experiment

The agreement between the empirical analysis[13] and current simulations is good, but obviously only qualitative due to the simplicity of the computational model used. Thus, while the mean experimental midpoint ϕ-value for over 800 mutations is ~0.36, we obtain a higher mean value of ~0.42–0.53 depending on the protein structural class and method used to calculate ϕ-values. The larger mean number is the result of using the mean-field Gō-model, where there is a direct relation between the stabilization energy and the degree of formed native structure (Fig. 6d). The average amount of stabilization energy gained in the TSE is roughly 50% for each of the studied proteins (Fig. 6d) agreeing well with the calculated ϕ-values. Accordingly, the experimental numbers would then correspond to ~36% of the stabilization energy gained at the barrier top under midpoint conditions (or a mean ϕ-value of 0.36[13]) and just ~24% gained at folding conditions (or a mean ϕ-value of 0.24[1,13]). This would suggest that the degree of native-like structure observed in our simulated TSE is actually an upper-limit of the experimental reality. In other words, the TSE of proteins is potentially even more unstructured and probably more diverse than that discussed in the last sections.

An alternate methodology is to constrain ensembles based on the experimental ϕ-values to obtain putative transition states.[64] The advantage of this methodology is that one eliminates any force-field or sampling issues that might influence the result. A recent comprehensive work on the protein acylphosphatase (α–β protein) sought to do precisely that, *i.e.* generate a constrained TSE based not only on ϕ-values but also that satisfies the p_{fold} criteria,[65] an aspect that was overlooked in the previous studies. The authors obtain a TSE that has a mean RMSD of ∼9 Å from the native state and conclude that it is quite diverse with a mean fraction of native contacts of 0.37 and lacking non-local contacts, results that strongly support our conclusions.

Future directions

The correlation of energetics with structure—an approximation of Gō-models—seems to be sufficient to reproduce the basic features of folding as we show here. This can be seen as evidence for the minimal frustration principle of the energy landscape theory[14] on which these models are based, wherein non-native interactions are assumed to have been evolutionarily weeded out resulting in a smooth folding landscape.[66,67] In terms of model development, it should be critical to be able to reproduce the general behaviour of protein folding systems (for example, the ϕ-value changes as a function of stability or the mean ϕ-value at midpoint conditions) first that can be extended further to incorporate additional features. It is of interest that a recent one-dimensional free energy surface model is able to reproduce this general behaviour, providing a simple and quantitative tool to analyze experimental data.[68,69] The Gō-model, on the other hand, has still room for improvement to be in better quantitative agreement with experimental numbers; probably by incorporating residue-specific energetic terms, attractive long-range terms, or a better dihedral description, issues that we are currently working on.

The general conclusions presented here and in the empirical work[13] are on the average behaviour and there are bound to be deviations, particularly when considering energetic effects like long-range electrostatics. A clear experimental test to the empirical observation that smaller destabilizations result in large ϕ-values is challenging as one has to deal with experimental noise when the destabilizations are less than 6–7 kJ mol^{-1} where the available dynamic range is quite small. This is complicated by the fact that protein surface residues are predominantly composed of electrostatic residues suggesting that the two-point methodology conventionally used in ϕ-values has potential limitations thus recommending the use of Davidson's multi-point ϕ-value analysis.[70] An even more exciting development is the identification of one-state downhill folding[71] and proteins that fold over marginal barriers.[72,73] They are characterized by a significant population of partially structured species even in equilibrium that can be studied at the atomic level by Nuclear Magnetic Resonance.[74] There are no transition states *per se* in these systems, but they still furnish a high-resolution picture of the species that are populated *en route* to folding (Fig. 7). Mutational effects on these proteins can be characterized at the same level of resolution under different stability conditions. The impact of mutational energetics on the partially structured species can then be followed directly providing valuable information on the interplay between hydrophobicity, packing requirements and electrostatics.

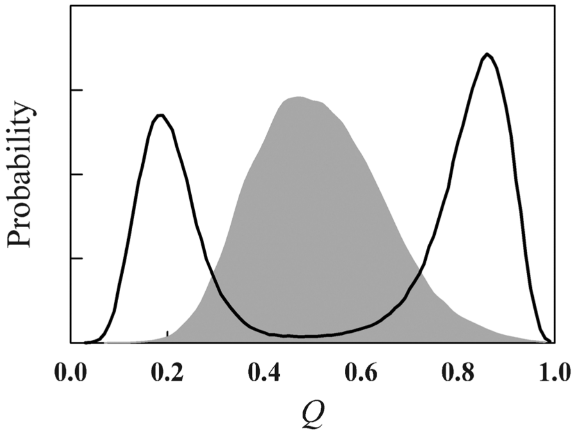

Fig. 7 Comparing one-state downhill and two-state protein folding. Comparison of the probability density at $T \approx 1$ for Ubq (black curve) with that of BBL (2CYU) at $T \approx 1.13$ (gray-filled area) as a function of the fraction of native contacts. The large probability density at intermediate reaction coordinate values for BBL has been probed experimentally by high-resolution NMR techniques.

Acknowledgements

This work was supported by the Juan de la Cierva Fellowship offered by the Spanish Ministry of Science and Innovation to A.N.N. and by research grants from the Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science), Instituto de Salud Carlos III (INB-Genoma España and COMBIOMED RETICS), Fundación Marcelino Botín and the European Union (EU)-ScalaLife project. The authors thank David de Sancho for valuable comments and critical reading of the manuscript.

References

1 A. Akmal and V. Muñoz, *Proteins*, 2004, **57**, 142–152.
2 A. N. Naganathan, U. Doshi, A. Fung, M. Sadqi and V. Muñoz, *Biochemistry*, 2006, **45**, 8466–8475.
3 M. Rueda, C. Ferrer-Costa, T. Meyer, A. Perez, J. Camps, A. Hospital, J. L. Gelpi and M. Orozco, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 796–801.
4 A. R. Fersht, A. Matouschek and L. Serrano, *J. Mol. Biol.*, 1992, **224**, 771–782.
5 L. S. Itzhaki, D. E. Otzen and A. R. Fersht, *J. Mol. Biol.*, 1995, **254**, 260–288.
6 D. P. Raleigh and K. W. Plaxco, *Protein Pept. Lett.*, 2005, **12**, 117–122.
7 I. E. Sanchez and T. Kiefhaber, *J. Mol. Biol.*, 2003, **334**, 1077–1085.
8 J. H. Cho and D. P. Raleigh, *J. Am. Chem. Soc.*, 2006, **128**, 16492–16493.
9 M. A. De Los Rios, B. K. Muralidhara, D. Wildes, T. R. Sosnick, S. Marqusee, P. Wittung-Stafshede, K. W. Plaxco and I. Ruczinski, *Protein Sci.*, 2006, **15**, 553–563.
10 J. N. Onuchic, N. D. Socci, Z. LutheySchulten and P. G. Wolynes, *Folding Des.*, 1996, **1**, 441–450.
11 H. S. Chan and K. A. Dill, *Proteins*, 1998, **30**, 2–33.
12 D. K. Klimov and D. Thirumalai, *Proteins*, 2001, **43**, 465–475.
13 A. N. Naganathan and V. Muñoz, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 8611–8616.
14 J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Proteins*, 1995, **21**, 167–195.
15 G. R. Bowman and V. S. Pande, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 10890–10895.
16 J. Zhang, W. F. Li, J. Wang, M. Qin and W. Wang, *Proteins*, 2008, **72**, 1038–1047.

17 J. W. Pitera, W. C. Swope and F. F. Abraham, *Biophys. J.*, 2008, **94**, 4837–4846.
18 F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich and T. R. Weikl, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 19011–19016.
19 C. D. Snow, E. J. Sorin, Y. M. Rhee and V. S. Pande, *Annu. Rev. Biophys. Biomol. Struct.*, 2005, **34**, 43–69.
20 H. Taketomi, Y. Ueda and N. Gō, *Int. J. Pept. Protein Res.*, 1975, **7**, 445–459.
21 B. A. Shoemaker, J. Wang and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 777–782.
22 M. Knott and H. S. Chan, *Proteins*, 2006, **65**, 373–391.
23 Y. Levy and J. N. Onuchic, *Acc. Chem. Res.*, 2006, **39**, 135–142.
24 Q. Lu and J. Wang, *J. Am. Chem. Soc.*, 2008, **130**, 4772–4783.
25 C. Clementi, H. Nymeyer and J. N. Onuchic, *J. Mol. Biol.*, 2000, **298**, 937–953.
26 C. Clementi, A. E. Garcia and J. N. Onuchic, *J. Mol. Biol.*, 2003, **326**, 933–954.
27 L. L. Chavez, J. N. Onuchic and C. Clementi, *J. Am. Chem. Soc.*, 2004, **126**, 8426–8432.
28 S. O. Garbuzynskiy, A. V. Finkelstein and O. V. Galzitskaya, *J. Mol. Biol.*, 2004, **336**, 509–525.
29 N. Koga and S. Takada, *J. Mol. Biol.*, 2001, **313**, 171–180.
30 J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu and J. N. Onuchic, *Nucleic Acids Res.*, 2010, **38**, W657–W661.
31 T. Veitshans, D. Klimov and D. Thirumalai, *Folding Des.*, 1997, **2**, 1–22.
32 B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.
33 H. Nymeyer, N. D. Socci and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 634–639.
34 S. S. Cho, Y. Levy and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 586–591.
35 R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka and E. S. Shakhnovich, *J. Chem. Phys.*, 1998, **108**, 334–350.
36 G. Hummer, *J. Chem. Phys.*, 2004, **120**, 516–523.
37 R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 6732–6737.
38 R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **107**, 1088–1093.
39 C. T. Friel, A. P. Capaldi and S. E. Radford, *J. Mol. Biol.*, 2003, **326**, 293–305.
40 H. M. Went and S. E. Jackson, *Protein Eng., Des. Sel.*, 2005, **18**, 229–237.
41 J. I. Guijarro, C. J. Morton, K. W. Plaxco, I. D. Campbell and C. M. Dobson, *J. Mol. Biol.*, 1998, **276**, 657–667.
42 B. Oztop, M. R. Ejtehadi and S. S. Plotkin, *Phys. Rev. Lett.*, 2004, **93**, 208105.
43 S. S. Cho, Y. Levy and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 434–439.
44 M. Oliveberg, Y. J. Tan, M. Silow and A. R. Fersht, *J. Mol. Biol.*, 1998, **277**, 933–943.
45 M. Jager, H. Nguyen, J. C. Crane, J. W. Kelly and M. Gruebele, *J. Mol. Biol.*, 2001, **311**, 373–393.
46 T. Ternstrom, U. Mayor, M. Akke and M. Oliveberg, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 14854–14859.
47 T. Y. Shen, C. P. Hofmann, M. Oliveberg and P. G. Wolynes, *Biochemistry*, 2005, **44**, 6433–6439.
48 D. E. Otzen, O. Kristensen, M. Proctor and M. Oliveberg, *Biochemistry*, 1999, **38**, 6499–6511.
49 G. Pappenberger, C. Saudan, M. Becker, A. E. Merbach and T. Kiefhaber, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 17–22.
50 M. Schatzle and T. Kiefhaber, *J. Mol. Biol.*, 2006, **357**, 655–664.
51 S. S. Plotkin, J. Wang and P. G. Wolynes, *J. Chem. Phys.*, 1997, **106**, 2932–2948.
52 I. E. Sanchez and T. Kiefhaber, *J. Mol. Biol.*, 2003, **327**, 867–884.
53 B. G. Wensley, M. Gartner, W. X. Choo, S. Batey and J. Clarke, *J. Mol. Biol.*, 2009, **390**, 1074–1085.
54 V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, *Biochemistry*, 1994, **33**, 10026–10036.
55 O. V. Galzitskaya and A. V. Finkelstein, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 11299–11304.
56 O. V. Galzitskaya, *J. Bioinf. Comput. Biol.*, 2008, **6**, 681–691.
57 O. V. Galzitskaya, *Biochemistry (Moscow)*, 2009, **74**, 186–193.
58 S. S. Plotkin and J. N. Onuchic, *J. Chem. Phys.*, 2002, **116**, 5263–5283.
59 Y. Suzuki and J. N. Onuchic, *J. Phys. Chem. B*, 2005, **109**, 16503–16510.
60 P. Weinkam, C. H. Zong and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 12401–12406.
61 C. A. Rohl, C. E. M. Strauss, K. M. S. Misura and D. Baker, *Methods Enzymol.*, 2004, **383**, 66–93.
62 V. Muñoz, *Curr. Opin. Struct. Biol.*, 2001, **11**, 212–216.
63 Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 2605–2610.
64 M. Vendruscolo, E. Paci, C. M. Dobson and M. Karplus, *Nature*, 2001, **409**, 641–645.
65 M. Kin, J. Zhang, H. Lu, R. Chen and J. Liang, *J. Chem. Phys.*, 2011, **134**, 075103.
66 J. E. Shea, J. N. Onuchic and C. L. Brooks, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 12512–12517.
67 J. E. Shea, J. N. Onuchic and C. L. Brooks, *J. Chem. Phys.*, 2000, **113**, 7663–7671.
68 A. N. Naganathan, U. Doshi and V. Muñoz, *J. Am. Chem. Soc.*, 2007, **129**, 5673–5682.
69 V. Muñoz, M. Sadqi, A. N. Naganathan and D. de Sancho, *HFSP J.*, 2008, **2**, 342–353.
70 J. G. B. Northey, K. L. Maxwell and A. R. Davidson, *J. Mol. Biol.*, 2002, **320**, 389–402.
71 M. M. Garcia-Mira, M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz and V. Muñoz, *Science*, 2002, **298**, 2191–2195.
72 A. Fung, P. Li, R. Godoy-Ruiz, J. M. Sanchez-Ruiz and V. Muñoz, *J. Am. Chem. Soc.*, 2008, **130**, 7489–7495.
73 A. N. Naganathan, P. Li, R. Perez-Jimenez, J. M. Sanchez-Ruiz and V. Muñoz, *J. Am. Chem. Soc.*, 2010, **132**, 11183–11190.
74 M. Sadqi, D. Fushman and V. Muñoz, *Nature*, 2006, **442**, 317–321.