PCCP

Cite this: *Phys. Chem. Chem. Phys.*, 2011, **13**, 17030–17043

www.rsc.org/pccp

PAPER

Integrated prediction of protein folding and unfolding rates from only size and structural class†

David De Sancho[ab] and Victor Muñoz*[ac]

Received 15th February 2011, Accepted 11th May 2011
DOI: 10.1039/c1cp20402e

Protein stability, folding and unfolding rates are all determined by the multidimensional folding free energy surface, which in turn is dictated by factors such as size, structure, and amino-acid sequence. Work over the last 15 years has highlighted the role of size and 3D structure in determining folding rates, resulting in many procedures for their prediction. In contrast, unfolding rates are thought to depend on sequence specifics and be much more difficult to predict. Here we introduce a minimalist physics-based model that computes one-dimensional folding free energy surfaces using the number of aminoacids (N) and the structural class (α-helical, all-β, or α–β) as only protein-specific input. In this model N sets the overall cost in conformational entropy and the net stabilization energy, whereas the structural class defines the partitioning of the stabilization energy between local and non-local interactions. To test its predictive power, we calibrated the model empirically and implemented it into an algorithm for the PREdiction of Folding and Unfolding Rates (PREFUR). We found that PREFUR predicts the absolute folding and unfolding rates of an experimental database of 52 proteins with accuracies of ±0.7 and ±1.4 orders of magnitude, respectively (relative to experimental spans of 6 and 8 orders of magnitude). Such prediction uncertainty for proteins vastly varying in size and structure is only two-fold larger than the differences in folding (±0.34) and unfolding rates (±0.7) caused by single-point mutations. Moreover, PREFUR predicts protein stability with an accuracy of ±6.3 kJ mol^{-1}, relative to the 5 kJ mol^{-1} average perturbation induced by single-point mutations. The remarkable performance of our simplistic model demonstrates that size and structural class are the major determinants of the folding landscapes of natural proteins, whereas sequence variability only provides the final 10–20% tuning. PREFUR is thus a powerful bioinformatic tool for the prediction of folding properties and analysis of experimental data.

Introduction

One of the important issues in protein folding is to understand how the chemical properties of proteins determine the rates of folding and unfolding, and from them the stability in native conditions. Traditionally, most efforts have concentrated on the folding rates. Theoretical arguments state that the native three-dimensional structure plays a critical role in folding, not only because it is the end point of the process, but also because it determines the funnelled shape of the energy landscape.[1,2] The curvature of the funnel dictates the magnitude of the entropic barrier, which in turn controls the rate of folding.[1] The importance of the 3D structure was first demonstrated empirically by the discovery of a correlation between folding rates and the contact order, a measure of the mean separation between residues within spatial contact in the native structure.[3] Immediately after, it was shown that Ising-like statistical mechanical models, in which the only possible interactions are native-like as prescribed by the Gō's consistency principle,[4] could predict the relative changes in rates with similar accuracy.[5] Moreover, the Ising-like models also reproduced the absolute span in experimental folding rates,[5] effectively connecting the contact order correlation with the physical mechanisms proposed by energy landscape theory. These two important results justified the use of native-centric modelling, which has since then spawned into hundreds of coarse-grained computer simulations of protein folding parameterized with the matrix of contacts derived from experimental 3D structures.[6,7] In the last years, efforts have focused on implementing native-centric computer models with more realistic descriptions of protein interactions. These new descriptions include statistical potentials for backbone propensities, non-native interactions, and/or non-additive terms.[8–10] Typically, these models have been tested against available experimental data on the changes in folding rates

[a] *Centro de Investigaciones Biológicas, Spanish National Research Council (CSIC), Ramiro de Maeztu 9, Madrid 28040, Spain*
[b] *Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK*
[c] *Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA. E-mail: vmunoz@cib.csic.es*

† This article was submitted as part of a themed collection on the Physical Foundations of Protein Folding.

induced by mutation (φ-values),[11,12] but it is not clear how well they reproduce experimental folding rates.

Moreover, structure is not the only determinant of folding rates. Scaling laws from polymer physics ascribe a critical role to protein size.[13,14] Based on such scaling principles, folding times should scale as $\log(\tau) \approx N^{\nu}$, in which N is the number of residues and the exponent could vary between 2/3 and nearly 0.[13–17] This overly simple equation does have, in fact, significant predictive power, as has been recently demonstrated on an experimental database of folding rates that included data for a wide range of protein sizes.[18] In parallel, other researchers have worked on developing novel structural metrics that increase the empirical correlation between folding rates and protein structure, whether by introducing size-scaling concepts,[19,20] refining the structural description,[21,22] considering differences between structural classes[23] or combining these ideas together.[24] Some of these new metrics are significantly better predictors than the contact order when applied to the improved experimental database that is available today (see Table 1), which spans 6 orders of magnitude in rates, sizes

Table 1 Protein dataset used in this study

	Protein	PDB ID	k_{f,H_2O} (s^{-1})	k_{u,H_2O}/s^{-1}	$[D]_0$/M	T_{exp}/K	pH	Struct. type	N
1	ABP1 SH3[40]	1jo8	1.17×10^{1}	6.59×10^{-2}	0.3	298	7	β	58
2	bACBP[40]	2abd	1.05×10^{3}	2.11×10^{-2}	0.5	298	7	α	86
3	ctAcp[62]	2acy	2.31	1.50×10^{-3}	0.2	301	5.5	α + β	98
4	mAcP[63]	1aps	2.30×10^{-1}	1.10×10^{-4}	0.2	301	5.5	α + β	98
5	ADA2h[64]	1o6x	7.58×10^{2}	4.82×10^{-1}	1.0	298	7	α + β	81
6	Azurin (apo)[40]	1e65	1.36×10^{2}	1.80×10^{-2}	0.4	298	7	α + β	128
7	BBL (H166W)[65]	2bth	1.30×10^{5}	6.00×10^{3}	2.0	298	7	α	45
8	BDPA[66]	1ss1	9.68×10^{4}	2.50×10^{1}	2.0	298	5.5	α	60
9	CheW[40]	1k0s	1.70×10^{3}	5.84×10^{-6}	2.2	298	7	α + β	151
10	CI2[67]	2ci2	4.78×10^{1}	1.81×10^{-4}	0.0	298	6.3	α + β	64
11	CspA[68]	1mjc	2.64×10^{2}	1.90×10	0.6	298	7	β	69
12	CspB Bc[69]	1c9o	1.37×10^{3}	6.40×10^{-1}	0.6	298	7	β	66
13	CspB Bs[70]	1csp	1.07×10^{3}	1.20×10^{1}	0.6	298	7	β	67
14	CspB Tm[69]	1g6p	5.65×10^{2}	1.80×10^{-2}	0.6	298	7	β	66
15	Cyt b562[a][71]	1yza	4.43×10^{3}	4.87×10^{-4}	2.0	298	5.2	α	106
16	E3BD (F166W)[65]	1w4e	2.75×10^{4}	2.00×10^{1}	0.7	298	5.5	α	45
17	EC298[40]	1ryk	8.78×10^{3}	8.91×10^{1}	0.2	298	7	α	69
18	Engrailed HD[56]	1enh	3.99×10^{4}	2.10×10^{3}	0	298	5.7	α	54
19	FBP28 (W30A)[51]	1e01	4.10×10^{4}	5.24×10^{3}	1.0	298	7	β	37
20	9 Fibronectin III[72]	1fnf	4.00×10^{-1}	5.27×10^{-2}	0.1	298	7.2	β	90
21	10 Fibronectin III[b][72]	1fnf	1.55×10^{2}	5.20×10^{-3}	0.7	298	7.2	β	94
22	FKBP12[73]	1fkb	4.30	1.70×10^{-4}	0	298	7.5	α + β	107
23	Fyn SH3[74]	1shf	9.43×10^{1}	9.90×10^{-4}	0.7	293	7.2	β	59
24	Hpr[75]	1poh	1.49×10^{1}	2.09×10^{-3}	0.5	293	7	α + β	85
25	Im7[40]	1ayi	1.34×10^{3}	1.04×10^{1}	0.8	298	7	α	86
26	Im9[40]	1imq	1.52×10^{3}	1.54×10^{-1}	1.0	298	7	α	86
27	L23[76]	1n88	2.04×10^{1}	7.41×10^{-5}	0.3	298	6.3	α + β	96
28	λ-repressor$_{6-85}$[40]	1lmb	3.22×10^{4}	2.48×10^{1}	1.0	298	8	α	80
29	cMyb[56]	1idy	6.20×10^{3}	5.30	0.3	298	5.7	α	54
30	PI3K SH3[77]	1pnj	3.53×10^{-1}	6.70×10^{-4}	0.15	293	7.2	β	86
31	POB (YWLA)[65]	1w4j	2.10×10^{5}	5.50×10^{2}	2.0	298	5.7	α	51
32	Protein G[78]	1pgb	4.12×10^{2}	1.28×10^{-1}	0.5	295	6	α + β	56
33	Protein L[79]	2ptl	6.06×10^{1}	2.00×10^{-2}	0.3	295	7	α + β	62
34	PTL9 C[80]	1div	2.63×10^{1}	3.90×10^{-4}	0.5	298	8	α + β	92
35	RafRBD[40]	1rfa	4.27×10^{3}	6.27×10^{-2}	1.0	298	7	α + β	78
36	hRAP1[56]	1fex	3.60×10^{3}	1.80×10^{1}	0.3	298	5.7	α	59
37	S6[81]	1ris	3.32×10^{2}	3.09×10^{-4}	0.4	298	6.2	α + β	97
38	Sho1 SH3[40]	2vkn	8.25	8.29×10^{-2}	0.3	298	7	β	66
39	α-Spectrin SH3[‡][82]	1shg	8.72	3.46×10^{-2}	0.3	298	3.5	β	57
40	Src SH3[83]	1rlq	5.67×10^{1}	1.00×10^{-1}	0.4	295	6	β	56
41	Src SH2[40]	1spr	6.25×10^{3}	3.08×10^{-2}	3.5	298	7	α + β	103
42	Sso7d (Y34W)[84]	1bf4	1.04×10^{3}	3.92×10^{-2}	1.2	293	6.1	α + β	63
43	Tenascin[85]	1ten	6.02	7.21×10^{-5}	0.2	298	7	β	90
44	Tendamistat[86]	3ait	6.66×10^{1}	4.50×10^{-5}	0.5	298	7	β	74
45	Tm1023[40]	1j5u	9.44×10^{2}	5.20×10^{-3}	2.0	298	7	α + β	125
46	hTRF1[56]	1ba5	3.70×10^{2}	3.20	0.3	298	5.7	α	53
47	Twitchin[c][87]	1wit	1.50	2.80×10^{-4}	0	293	7	β	93
48	U1A[d][88]	1urn	3.78×10^{1}	2.46×10^{-2}	0	298	6.3	α + β	96
49	Ubiquitin[40]	1ubq	1.52×10^{3}	1.07×10^{-3}	0.3	298	5	α + β	76
50	Urm1[40]	2qjl	1.32×10^{1}	3.69×10^{-2}	0.5	298	7	α + β	99
51	WW prototype[51]	1e0m	7.00×10^{3}	1.20×10^{3}	0	298	7	β	37
52	Yap[51,65]	1k9q	4.30×10^{3}	7.80×10^{2}	0	298	7	β	40

[a] The folding and unfolding rate constants correspond to 2.5 M urea, as reported originally by the authors. [b] Only refolding limb was originally reported. k_u and m_u were calculated using ΔG_{eq} from the equilibrium experiment. [c] The values shown correspond to our own fit to a two state model of the digitized experimental data. [d] Due to the pronounced curvature of the chevron plot at low concentrations of denaturant we use the folding and unfolding rate constants at 2 M GdHCl instead of those in water.

Table 2 Correlation coefficients of structural parameters and folding rates in water with confidence intervals from bootstrap analysis

	R_{H_2O}	CI_{H_2O}
$N^{1/2}$	−0.42	(−0.61, −0.17)
RCO	−0.68	(−0.78, −0.53)
ACO	−0.72	(−0.84, −0.54)
% Local	0.62	(0.46, 0.74)
LRO	−0.78	(−0.85, −0.66)
TCD	−0.76	(−0.85, −0.64)
L_{eff}	−0.66	(−0.76, −0.28)
Q_{d}	−0.75	(−0.85, −0.57)

ranging from 30 to 150 residues, and eliminates proteins with large prosthetic groups and/or multistate kinetics. The prediction performance of the various structural descriptors on this database is shown in Table 2.

Much less is known about what determines the barrier crossings in the other direction: the unfolding rates. This is in part because the prediction of unfolding rates appears to be much more challenging. In the first attempts, experimental unfolding rates did not show any apparent trends when correlated against simple structural features, size, or the native state stability.[25] A posterior bioinformatics effort has been successful in reproducing unfolding rates for 26 proteins by combining a large array of properties from amino acid sequences into an empirical equation with adjustable weights.[26] Very recently, unfolding rates predicted from a free energy surface model have been shown to correlate well with experiment,[27] although not in absolute terms since the predicted dynamic range in rates was grossly overestimated. In parallel, on a previous study using a minimalist model that represents folding as diffusion on a one-dimensional free energy surface we showed that the whole experimental variability in folding rates and stabilities of 52 proteins arises from minimal fluctuations in the two basic model parameters.[28] The stability data could be reproduced exactly with only 1.7% variability in the net stabilization energy per residue, whereas folding rates required 8% changes in the parameter that determines the barrier height.[28] These results suggested that the changes in folding and unfolding rates compensate each other to a large extent, and thus that folding and unfolding are determined by the same physico-chemical factors.

Thus a critical target for further studies is to achieve an integrated prediction of absolute folding and unfolding rates (and by extension of protein stability) that incorporates structural and size-scaling factors into a physics-based theoretical model. Such model would permit to assess the relative roles of size and structure in (un)folding, serve as tool for the prediction and interpretation of experimental data, and provide the foundations for future implementation of sequence-dependent information. Here we address this issue making use of a minimalist one-dimensional (1D) free energy surface model.[29] This model represents the folding energy landscape as a mean-field projection onto a single local order parameter termed *nativeness*.[29] Folding–unfolding kinetics are then obtained as diffusion on this mean-field 1D free energy surface.[30] The previous model implementation, which only accounted for size effects, reproduces the empirical size-scaling observed for entropy, enthalpy and heat capacity of unfolding,[31] folding rates[18] and protein stability.[28] In this work we introduce protein structure into the model by separating the stabilization energy into two net contributions: from local interactions (between residues close in sequence) and from long-range interactions. This description is in the same spirit of early lattice simulations,[22,32] Ising-like models,[5,33] experiments that engineer secondary structure propensities,[34] and structural analysis.[35] We calibrate the local and non-local energy contributions of the model from: (1) an analysis of an experimental database of 52 proteins (Table 1), and (2) the comparison between such empirical energies and the number of contacts observed in protein 3D structures. Finally, we show that such physics-based model simultaneously predicts absolute folding and unfolding rates using size (*N*) and ascription to one of the three structural-classes (all-α, α–β, all-β) as only protein-specific information. The prediction accuracy (correlation coefficient) for the absolute folding rates is as good as that of the best prediction of relative rates from structural metrics (*LRO*).[21] More remarkably, the performance on the unfolding rates is also reasonable. Thus suggesting that, contrary to what permeated from early studies,[25] a simple integrated-quantitative theory for predicting protein folding rates, unfolding rates and stability, is feasible.

The outline of the article is as follows. First we describe the general features of the mean-field free-energy surface model and introduce the procedure to account for local and non-local energies. In the next two sections we apply the model to estimate the local and non-local contributions for the 52 proteins in our database and perform a simple empirical test of the significance of these energies. At this point we also discuss some interesting findings about the combined roles of structure and sequence in determining folding energy landscapes that emerge from our analysis. In the fourth section we exploit the results from the model to perform a comparative analysis between folding energetics and structure. We investigate how well the matrices of native contacts from 3D structures compare with the empirical local and non-local stabilization energies; and we present a simple procedure to convert structural contacts into local and non-local stabilization energies that is structural-class specific. Finally, we describe the implementation of these findings into the algorithm PREFUR, which either predicts the 1D folding free energy surface of a protein (and thus the absolute folding and unfolding rates) using its size and structural class as only input, or calculates the mean content in local and non-local interactions of a protein from the known experimental folding and unfolding rates (available as a webserver application at: http://tmg.cib.csic.es/servers/PREFUR).

Methods

Free energy surface model

As starting point we have used the one dimensional free energy surface model previously developed in our laboratory.[29] This model is based on a mean-field approximation, and thus it should only be used for proteins that are structurally homogeneous (single-domain proteins). The model was first employed to rationalize the observation of systematic deviations from two-state folding behavior in the chemical and thermal unfolding kinetics of ultra-fast folding proteins (*i.e.* with microsecond folding-times).[29] The model has also been used to investigate size-scaling effects to protein folding stability and rates[28] and for the quantitative analysis of the folding–unfolding experiments of the following proteins: gpW,[36] BBL[37] and its structural homolog PDD,[38] and monomeric λ-repressor.[39]

In the model the energy landscape is represented as a one-dimensional free energy surface resulting from the projection onto a local order parameter termed nativeness (n). The value of nativeness corresponds to the average probability of finding any residue in the protein populating native-like φ–ψ dihedral angles, and thus it ranges from 0 to 1. Therefore, n is a continuous analogue of the fraction of residues in native conformation. For simplicity, the conformational space of a residue is divided into two states: the folded state, which corresponds to a small area of the Ramachandran map centered about the dihedral angles observed in the native 3D structure; and the fully-unfolded state that accounts for the remainder of the sterically allowed region of the map. Defining both residue-states as microcanonical ensembles, the conformational entropy of a residue in the fully-unfolded and native states are $S_{\mathrm{conf,res}}(n = 0) = R\ln(\Omega_{\mathrm{U}})$ and $S_{\mathrm{conf,res}}(n = 1) = R\ln(\Omega_{\mathrm{F}})$, respectively. Ω_{U} and Ω_{F} are the fully-unfolded and folded regions of the Ramachandran map. The conformational entropy of the residue at any other value of n is

$$S_{\mathrm{conf,res}}(n) = -R[n\ln(n) + (1-n)\ln(1-n)] + R[n\ln(\Omega_{\mathrm{F}}) + (1-n)\ln(\Omega_{\mathrm{U}})] \qquad (1)$$

where the first term reflects the combinatorial entropy and the second accounts for the intrinsic difference in entropy between the native and fully-unfolded states. Eqn (1) is obtained calculating the entropy of the macrocanonical ensemble defined by the mix of residues in folded and unfolded conformation using the Gibbs entropy formula, and noting that at a given value of n the probability of the residue being on each of the microstates belonging to the folded ensemble is n/Ω_{F}, and the probability of being on one of the microstates belonging to the fully unfolded ensemble is $(1-n)/\Omega_{\mathrm{U}}$. From eqn (1) and defining the folded state ($n = 1$) as reference, we obtain the previously used expression for the changes in conformational entropy as a function of n for a protein of N residues[29]

$$\Delta S_{\mathrm{conf}}(n) = N(-R[n\ln(n) + (1-n)\ln(1-n)] + (1-n)\Delta S_{\mathrm{conf,res}}) \qquad (2)$$

where $\Delta S_{\mathrm{conf,res}} = R\ln(\Omega_{\mathrm{U}}/\Omega_{\mathrm{F}})$ is a basic parameter of the model that defines the difference in conformational entropy between the fully-unfolded and folded states of a protein residue (not to confuse with the cost in conformational entropy upon folding which is determined by the difference between the two minima in the free energy surface).

The changes in stabilization energy (enthalpy) as a function of nativeness are modeled as a Markov-chain. That is, the probability of breaking existing interactions upon unfolding is assumed to be constant and proportional to differential changes in nativeness (δn), resulting in an exponential decay. In the original formulation of the model the stabilization energy was described with the phenomenological exponential function:[29]

$$\Delta H(n) = N\Delta H_{\mathrm{res}}[1 + (\exp(\kappa_{\Delta H}n) - 1)/(1 - \exp(\kappa_{\Delta H}))] \qquad (3)$$

The sharpness of the exponential decay for the mean-field stabilization energy is thus an adjustable parameter ($\kappa_{\Delta H}$) that determines the shape of the resulting 1D free energy surface, which can range from two-state with a high barrier to one-state downhill.[29] However, it is important to note that there is an equivalent equation that can be obtained in probabilistic terms from the Markov model:

$$\Delta H(n) = N\Delta H_{\mathrm{res}}[(1 - x^{(1-n)})/(1 - x)] \qquad (4)$$

where x determines the characteristic rate of breaking native interactions for infinitesimal changes in nativeness, and $(1 - x^{(1-n)})/(1 - x)$ is the fraction of remaining native stabilization energy as a function of n (0 for $n = 0$ and 1 for $n = 1$). Eqn (3) and (4) are interchangeable by noting that $\kappa_{\Delta H} = -\ln(x)$. From here onwards we will use eqn (4) because it is simpler and more rigorous. Regarding the characteristic parameter, x is always positive, but $x < 1$ implies that most of the stabilization energy is realized at high n values (*i.e.* the function is convex), whereas $x > 1$ implies that a large fraction of the stabilization energy is realized already at low values of n (*i.e.* concave function) Therefore, one can separate the stabilization energy into two terms with different characteristic curvature to represent the contributions from local and non-local interactions:

$$\Delta H(n) = N(\Delta H_{\mathrm{res,loc}}[(1 - x_{\mathrm{loc}}^{(1-n)})/(1 - x_{\mathrm{loc}})] + \Delta H_{\mathrm{res,non\text{-}loc}}[(1 - x_{\mathrm{non\text{-}loc}}^{(1-n)})/(1 - x_{\mathrm{non\text{-}loc}})]) \qquad (5)$$

The advantage of this formulation is that it accounts explicitly for the fact that making local interactions requires fixing much fewer degrees of freedom than tertiary interactions. In this case the specific curvatures of the local and non-local interactions terms are constant and what varies from protein to protein are the per-residue magnitudes of the two contributions (ΔH_{loc} and $\Delta H_{\mathrm{non\text{-}loc}}$).

Likewise, the change in heat capacity, which was introduced before also as a Markov-chain, can be expressed as

$$\Delta C_p(n) = N\Delta C_{p,\mathrm{res}}[(1 - x_c^{(1-n)})/(1 - x_c)] \qquad (6)$$

where $\Delta C_{p,\mathrm{res}}$ is the per-residue difference in heat capacity between the fully-unfolded state ($n = 0$) and the native state ($n = 1$) and x_c the characteristic rate of change for the heat capacity [$x_c = \exp(-\kappa_{\Delta Cp})$].

Using eqn (2), (5) and (6) and 385 K as T_{ref} (temperature at which the solvation free energy cancels out),[31] the free energy surface at any given temperature is obtained as:

$$\Delta G(n) = \Delta H(n) - T\Delta S(n) + \Delta C_p(n) \times [(T - T_{\mathrm{ref}}) + T\ln(T/T_{\mathrm{ref}})] \qquad (7)$$

The balance between the enthalpy and the entropy functions determines protein stability (ΔG_{UF}), whereas the difference in shape between the functionals for the entropy and enthalpy (determined by the curvature of the total stabilization energy, which in turn depends on the relative contributions from local and non-local interactions) produces a free energy barrier ($\beta_{U\ddagger}$) at values of n between 0.7 and 0.8. The free energy surface determines folding–unfolding thermodynamics directly. Folding kinetics can be calculated as diffusion on the same surface. For simplicity, here we calculate folding and unfolding rates directly from the forward and backward barrier heights assuming an appropriate pre-exponential term: $k = k_o \exp(-\beta/RT)$. For proteins with barriers above $\sim 2RT$ this calculation renders essentially the same results than the accurate diffusion treatment. Hence the resulting model has 8 parameters $\Delta S_{\text{conf,res}}$, $\Delta H_{\text{res,non-loc}}$, $x_{\text{non-loc}}$, $\Delta C_{p,\text{res,xc}}$ and k_o, most of which we fix here to empirical values from a two-state analysis of thermodynamic data[31] (see below).

Protein kinetics database

In this work we use experimental kinetic data (*i.e.* relaxation rates as a function of the concentration of denaturant, or chevron plots) for 52 proteins. The kinetic data for all these proteins did not exhibit any signs of transient population of folding intermediates and were obtained in similar experimental conditions (298 ± 5 K and nearly neutral pH) (Table 1). The advantage of such database is that it eliminates the complexity of multi-state folding and the intrinsic effects of temperature on folding kinetics (changes in the pre-exponential or diffusion coefficient), which are very strong.[29] This database thus complies with the recommendations on the 'standard' set of experimental conditions for quantitative biophysical studies of protein folding.[40] For all the proteins in the dataset we show the folding and unfolding rate constants extrapolated to water conditions ($k_{\text{f,H}_2\text{O}}$ and $k_{\text{u,H}_2\text{O}}$), the lowest denaturant concentration measured experimentally ($[D]_0$), experimental temperature (T_{exp}) and pH (Table 1).

Structural analysis

The experimental 3D structures of the proteins in our database were downloaded from the RCSB Protein Data Bank.[41] We edited the coordinate files removing residues from the tails to achieve maximum agreement with the protein constructs actually used for the kinetic experiments whenever a description of the construct was provided in the original work. For comparison with our predictions, we show the values for some of the most successful structural metrics for prediction of folding rates reported in the literature: relative contact order (*RCO*),[3] absolute contact order (*ACO*),[20] fraction of local contacts (% *local*),[22] long range order (*LRO*),[35] total contact distance (*TCD*),[24] number of sequence-distant native pairs (Q_d),[42] and effective length ($L_{\text{eff}}{}^{P}$, with $P = 0.1$)[43] (see Table 2). Our definition of native contacts is similar to that used before by other authors.[9,44] Two amino acid residues i and j form a native contact if any of their inter-atomic distances is shorter than 5 Å. Contacts are considered local when the sequence separation is shorter than or equal to 3 residues while otherwise they are counted as non-local.

Kinetic data fits, correlation analysis and rate predictions

To obtain perfect fits of the model to individual proteins, we adjusted the two parameters defining the local and non-local stabilization energies per residue ($\Delta H_{\text{res,loc}}$ and $\Delta H_{\text{res,non-loc}}$) to reproduce the experimental folding and unfolding rates exactly using a simplex minimization method. We performed correlation analysis and bootstrap tests to obtain 95% confidence intervals for the correlation coefficients. All calculations were carried out using MATLAB (The MathWorks) software package, together with StarP when the calculations were run in parallel.

Bioinformatics tools

For the comparison of protein 3D structures of the same superfamilies we used the Secondary Structure Matching tool from the European Bioinformatics Institute, authored by E. Krissinel and K. Henrick[45] (http://www.ebi.ac.uk/msd-srv/ssm). For the prediction of the helical content of proteins we used the AGADIR server at: http://agadir.crg.es

Results and discussion

Estimating the contributions from local and non-local interactions to protein folding

Since the early equilibrium statistical mechanical models for protein folding, it has been common practice to separate the energetic contributions to protein folding into two major terms, one corresponding to interactions between residues close in sequence (local interactions) and another accounting for interactions between residues further away (non-local interactions).[5,32,35,46] The rationale is that local interactions form upon fixing very few degrees of freedom in the polypeptide chain and thus involve small entropy penalties that are compatible with highly unfolded conformations. This concept was reinforced by the successful predictions of folding rates by the contact order[3] and simple Ising-like models,[5] which suggest that local interactions form earlier during folding and speed up the folding rate, whereas non-local interactions slow down folding. A similar conclusion has been drawn more recently from the large-scale analysis of mutational data in protein folding.[47]

Introducing a distinction between net energetic contributions from local and non-local interactions to the free energy surface model is, in principle, straightforward. We achieve this by separating the stabilization energy into two additive contributions: local and non-local. Each of these contributions is characterized by two specific parameters defining the net contribution per residue (ΔH) and the curvature of the dependence on nativeness (x) (see eqn (5)). The issue is how to parameterize the specific dependence of local and non-local interactions on the order parameter nativeness (*i.e.* the parameter x). For the sake of discussion it is convenient to address this issue by comparing the mean-field model with the classical Ising-like statistical mechanical model of folding.[5] These two models have the same treatment of the conformational entropy, to the extent that the entropy function in eqn (2) is the equivalent in the continuum of the conformational entropy arising from the full enumeration of microstates of the Ising-like model. The differences between models are found in the treatment of the stabilization energy. In the Ising model interactions are only made when all the peptide bonds connecting the interacting residues are simultaneously fixed in native conformation.[5] This restrictive rule enforces a mechanism in which folding nucleates locally and progresses by growth from the nuclei. The mean-field model defines the stabilization energy at a given value of n as the weighted average over all possible conformations compatible with such value. The lack of an explicit treatment for the interactions makes the connection with the protein native structure less explicit. In turn, the practical advantage is more mechanistic flexibility. For example, if we enforce the Ising rule to make interactions, the formation of local interactions (*e.g.* i, $i + 3$) would correspond in the mean-field model to a stabilization energy function that decays as n^3 (three adjacent residues simultaneously native). Such dependence can be approximated by eqn (4) using $x \approx 0.05$ (or $\kappa \approx 3$). The curvature of this local energy is, however, as steep as that previously obtained for the global stabilization energy (local plus non-local) of the previous version of the mean-field model.[28] Therefore, our version of the mean-field model employs *de facto* a rule for making native interactions that is much less restrictive than the classical Ising rule. That is, native interactions are allowed to form to certain degree without requiring all the connecting residues simultaneously fixed in native conformation. It follows that the curvature of the local term in eqn (5) should be much less steep than the average global term employed before ($x \gg 0.05$), whereas the non-local term should be steeper ($x < 0.05$). Another important point to consider is that the differences in the global parameter x required by the mean-field model to reproduce exactly the experimental folding rates of 52 proteins are of only 25% (8% in $\kappa_{\Delta H}$),[28] whereas the Ising rule implies drastic differences in the curvature of local and non-local interactions (*i.e.* a stabilization energy functional that decays as n^p, where p is the number of residues connecting the two residues involved in the interaction).

These considerations set some useful restrictions in the range of physically reasonable values to employ for the curvature of the local and non-local energy terms. To further evaluate this range we performed exact fits of the model to the experimental dataset of 52 folding and unfolding rates in aqueous solution (k_f and k_u), which we derived from a standard two-state analysis of the experimental relaxation rate and equilibrium constant (Table 1). In these exact fits we floated two model parameters for each protein, the specific values of ΔH_{loc} and $\Delta H_{non\text{-}loc}$, using a fixed two-dimensional grid of values for x_{loc} and $x_{non\text{-}loc}$. For consistency, we fixed the remaining model parameters to the empirical values estimated from our previous analysis: $\Delta S_{res} = 16.5$ J mol^{-1} K^{-1}, $\Delta C_{p,res} = 0.058$ kJ mol^{-1} K^{-1}, $x_c = 0.0136$ (or $\kappa_{\Delta Cp} = 4.3$), and a pre-exponential $k_o = 3.5 \times 10^6/N$ s^{-1} that is equivalent to using a diffusion coefficient $D = 8 \times 10^4\ n^2/N$ s^{-1} in the diffusive kinetic calculation.[28] We note that the values for the thermodynamic parameters ΔS_{res} and $\Delta C_{p,res}$ are derived from correlation analysis of calorimetric data,[31] and that the same pre-exponential is used in both folding and unfolding directions.

The results of these calculations are shown in Fig. 1, where the two-dimensional grid of values for the curvature of local and non-local energies is shown as $-\ln(x)$ to linearize the plot. This exercise reveals that there is a broad region of the grid, defined by $-\ln(x_{loc})$ between -1.5 ($x_{loc} = 4.5$) and 1.5 ($x_{loc} = 0.22$) and $-\ln(x_{non\text{-}loc})$ between 3.5 and 4 ($x_{non\text{-}loc}$ between 0.03 and 0.018), where the fits converge to the experimental values with negligible sums of least squares (*SLS*) (Fig. 1A). For the non-local

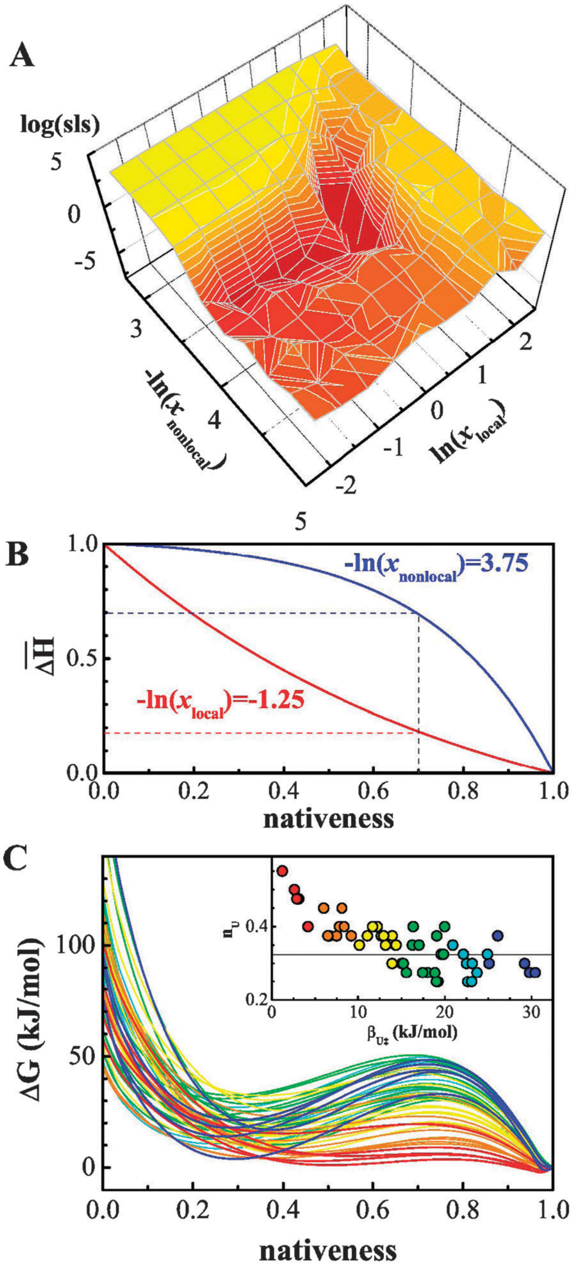

Fig. 1 Exploration of the parameter space for local and non-local stabilization energies. (A) Quality of the fits to the rates expressed as sum of least squares of experimental and calculated rates (SLS) *versus* the negative logarithm of the curvature of the stabilization energy functional for the local and non-local contributions (x_{loc} and $x_{non\text{-}loc}$). (B) The normalized functionals defining the changes in the local (red) and non-local (blue) stabilization energies ($\Delta\bar{H}$) as a function of the order parameter nativeness for the values of choice of x_{loc} and $x_{non\text{-}loc}$. (C) One dimensional free energy profiles (ΔG) as a function of nativeness resulting from the exact fit of the model to the experimental data of 52 proteins. Curves are colored according to the light spectrum so that the energy is proportional to the barrier height ($\beta_{U\ddagger}$). The inset shows the position of the unfolded state (n_u) as a function of $\beta_{U\ddagger}$. The gray line signals the average value of n_u for proteins with $\beta_{U\ddagger} > 10$ kJ mol^{-1} ($\sim 4\ RT$).

term the area covering the good combinations of parameters delimits a narrow range of curvatures that are slightly steeper than the global value. The local parameter varies more, producing local energy functions that range from slightly convex to slightly concave, signifying that a large fraction of local interactions is already present at the free energy surface minimum corresponding to the protein unfolded state. From all the mathematically equivalent solutions delimited by the low *SLS* area of Fig. 1A, we arbitrarily selected one that maximizes the difference in curvature between the local and non-local terms (x_{loc} = 4.5 and $x_{\text{non-loc}}$ = 0.0235; Fig. 1B). The rationale behind this decision is that by increasing the difference between local and non-local curvatures we make the model maximally responsive to the variability of protein 3D structures. From a theoretical viewpoint, the functional for local interactions shown in Fig. 1B is similar to assuming that making these interactions involves entropy costs similar to those of closing disordered loops in polypeptides, for which a scaling of $\sim n^{2/3}$ is expected.[12] As a practical advantage, this choice of parameters gives the model more flexibility to account for the connection between local interactions and experimental ϕ-values discovered recently.[47]

The exact fits to the experimental data using these local and non-local description produce free energy surfaces with the native minimum at very high values of nativeness ($n \approx 0.97$) separated of the unfolded minimum by a barrier of variable height located at n ranging from 0.7 to 0.8 (Fig. 1C). The position of the unfolded minimum shows much more variability. In fact, there is a strong correspondence between the position of the unfolded state and the height of the free energy barrier in the folding direction ($\beta_{U\neq}$) (see inset in Fig. 1C). For proteins with large folding barriers ($\beta_{U\neq} > 4RT$) the unfolded state is found at low values of the order parameter (n_U = 0.32 on average), with little variability from protein to protein. However for proteins with small or marginal barriers ($\beta_{U\neq} < 4RT$), the position of the unfolded state shifts towards higher values of the order parameter proportionally to the inverse of the barrier height. The properties of the free energy surfaces generated by the exact fits to the mean-field model can be summarized in two general predictions: (1) the height of the folding barrier should be proportional to the relative contribution from non-local interactions to the stabilization energy; (2) the unfolded state of proteins with marginal barriers is by necessity more native-like, including large fractions of local interactions. These predictions are consistent with the common finding of residual native-like backbone conformation in unfolded states of ultrafast folding proteins.[48–50]

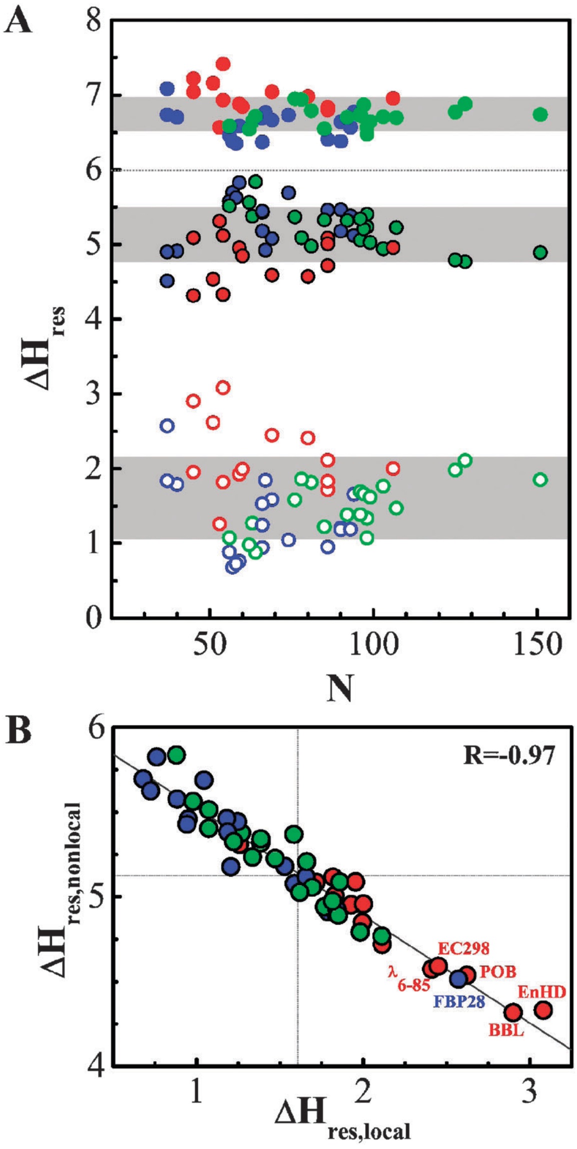

Fig. 2 Stabilization energy values resulting from the fit of folding rates. (A) The local stabilization energy per residue is shown as empty circles and the non-local as full circles with a black rim. The sum of the two terms (total ΔH per residue) is shown on top as full circles. The color code signifies the structural class: (blue) β-proteins, (green) α + β proteins, (red) α-helical proteins. (B) Correlation between the local and non-local stabilization energies per residue. All relevant quantities are in kJ mol^{-1}.

An almost perfect compensation between local and non-local contributions

The exact fits of the model to the experimental folding and unfolding rates allow us to estimate how much of the total stabilization energy of each protein corresponds to the local (ΔH_{loc}) and non-local ($\Delta H_{\text{non-loc}}$) contributions. The values for these two parameters obtained from the exact fits are shown in Fig. 2. The first observation from this exercise is that the total stabilization energy per residue is nearly constant for the 52 proteins (top in Fig. 2A). Therefore, the new implementation reproduces the result we obtained before with a simpler version of the model,[28] as well as the linear scaling of ΔH with protein size that has been observed in calorimetric data.[31] Inspection of the two separate contributions reveals that, on average, local interactions account for about one fourth of the stabilization energy of the protein. It is important to note, however, that this ratio is pre-determined by the relative curvatures of the local and non-local stabilization energies used in the model (Fig. 1B). On the other hand, the differences between individual proteins should be model independent. In contrast to the nearly constant total stabilization energy, the contributions from local and non-local interactions vary quite significantly (Fig. 2A). The fluctuations on the local stabilization energy per residue measured as the standard deviation are 0.55 kJ mol^{-1} res^{-1} (~34% of the mean), somewhat larger than the fluctuations of the non-local contribution, which are of only 0.36 kJ mol^{-1} res^{-1}. There are also hints of some trends in the data. For example, the largest fluctuations from the mean behavior are found for the smallest proteins (the left corner in Fig. 2A). In addition, the same small proteins appear to have lower non-local/local ratios than the rest. These two observations could be the result of capillarity effects by which the unfavorable volume/surface ratio of proteins with $N \lesssim 50$ results in very small hydrophobic cores.[15]

There are additional trends at the coarse structural level. Particularly, α-helical proteins exhibit larger local contributions than the average, whereas the opposite is true for β-proteins. The α + β proteins are closer to the average values. In this respect it is noteworthy that the few β-proteins of very small size included in our database (*i.e.* the WW domains FBP28, Yap and WW prototype)[51] are clear outliers of the general structural trend. This might be due to a dominance of size over structural effects, or related to the fact that these three β-proteins are all microsecond folders and candidates for downhill folding.[29,52] FBP28, in particular, exhibits local stabilization energies per residue that are well above the standard deviation of the entire dataset, consistently with the marginal thermodynamic cooperativity of this domain according to the parameter n_σ.[18,52]

The most intriguing discovery emerging from this analysis is the fact that there seems to be an almost perfect compensation between the local and non-local contributions to the stability of the 52 proteins included in the database. Such energetic compensation results in a very high anti-correlation ($R = -0.97$) between the local and non-local stabilization energies per residue (Fig. 2B). This is an important result that helps to explain our previous observation that the total stabilization energy per residue in this same dataset varies by a mere 1.7% from protein to protein.[28] We note that the local *versus* non-local compensation is also maintained on the proteins that are farthest away from the mean behavior (Fig. 2B). This is particularly evident for the subgroup of microsecond folders, which correspond to most of the outliers in Fig. 2A (low $\Delta H_{\text{res,non-loc}}$ and large $\Delta H_{\text{res,loc}}$), but are perfectly in trend in the anti-correlation shown in Fig. 2B. This group includes proteins of small size and/or characteristically weak hydrophobic cores, such as the peripheral subunit binding domains (PSBDs) BBL and POB, engrailed homeodomain, and the FBP28 WW domain. As part of this group we also find the λ-repressor$_{6-85}$, a protein that has later been engineered to fold downhill, and EC298, a large α-helical protein that folds in about 100 μs according to linear extrapolation.[40] However, it is noteworthy that other proteins from the PSBD, homeodomain like and WW domain families are closer to the average behavior (marked by the cross in Fig. 2B), indicating that the unusually high content in local interactions of the proteins of the lower right corner of Fig. 2B is a sequence-related property selected by evolution.

Empirical test of the model and analysis on two α-helical protein superfamilies

Our analysis is based on the premise that the modeled one-dimensional free energy surfaces lead to adequate calculations of folding kinetics, or, in other words, that the local order parameter nativeness is a reasonable reaction coordinate.[30,54] The results may also be affected by the shape of the local and non-local energy functionals (*e.g.* Fig. 2B) and choice of parameters. Regarding the latter, the specifics of the local and non-local energies and the magnitude of the conformational entropy determine the global partitioning between local and non-local contributions to the total energy, but do not change the relative differences between proteins. Their effects are thus equivalent to introducing a simple rescaling of the two energies, which does not significantly affect our analysis. But, it is important to test whether the differences in local (or non-local) contributions that we derive from the experimental folding–unfolding rates reflect the true energetic balance of the proteins under study. Such test can be performed for the local interactions of α-helical proteins by comparing our energies with the predictions from available helix-coil statistical mechanical models that accurately predict α-helix propensities from the aminoacid sequence alone.[55] To factor size and structural effects out, it is also preferable to compare proteins within superfamilies.

The experimental database of 52 proteins does include several members of two different α-helical superfamilies that could be used in this test: 4 members of the homeodomain-like proteins and 3 members of the peripheral subunit-binding domains (PSBD) of 2-oxo acid dehydrogenase complexes. The sequence homology of the available members of both, homeodomain-like and PSBDs, is actually low (Fig. 3A). Moreover, there is quite significant variability in folding–unfolding rates within the members of the two families in spite of sharing size and 3D structure. In Fig. 3B we show the superimposed 3D structures of the members of the two superfamilies.

The sequence homology of the four homeodomains (EnHD, hTRF1, hRAP and c-Myb) is always lower than 26%, but their structures share a three helix bundle fold with a characteristic helix-turn-motif and can be superimposed with an overall *RMSD* of 2.4 Å[56] (blue tones in Fig. 3B). Also the number of local (88, 86, 90, 95, respectively) and non-local contacts (114, 104, 93 and 115, respectively) is very similar for the four structures. However, our analysis shows that these proteins are very different energetically, as it could be expected given that their folding rates in water span three orders of magnitude.[56] EnHD has much more local stabilization energy (166.3 kJ mol^{-1}) than the other three proteins (hTRF1, 66.6 kJ mol^{-1}; c-Myb, 98.2 kJ mol^{-1}; RAP1, 113.7 kJ mol^{-1}), whereas the opposite holds for the non-local stabilization energy. Likewise, these proteins have helical propensities calculated from their aminoacid sequence using AGADIR[55] that differ by up to 10-fold. The local stabilization energies from our analysis do in fact correlate with the AGADIR helix propensities expressed in logarithmic scale (Fig. 3C). Similarly, the three PSBDs in our database (BBL, E3BD and POB) can be easily superimposed with an average *RMSD* of 1.3 Å (Fig. 3B). The sequences of BBL and POB are more closely related with one another (52%) than they are with E3BD (<40%). Our analysis indicates that E3BD has a much lower contribution from local interactions than BBL and POB (87.9 kJ mol^{-1} *versus* 130.5 and 133.6 kJ mol^{-1}).

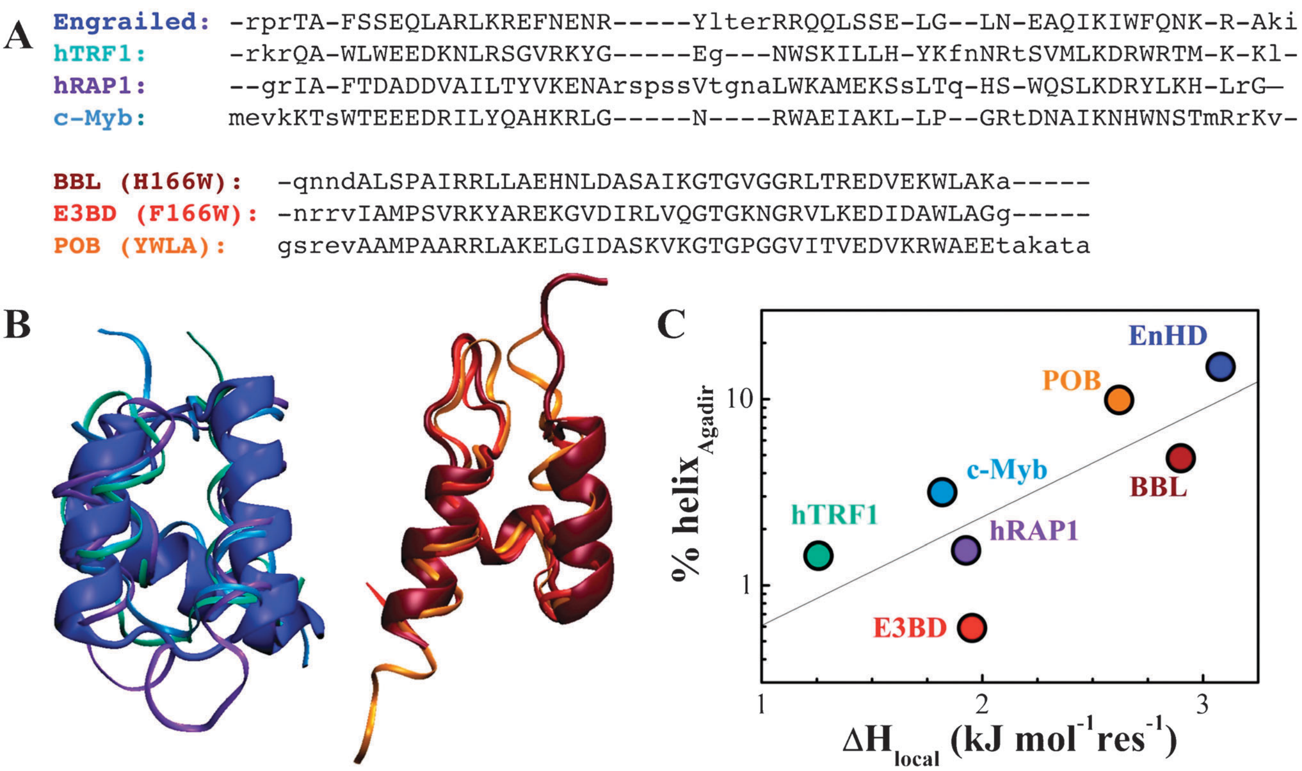

Fig. 3 Comparing local energy contributions and helix propensities in homeodomain-like proteins and peripheral subunit binding domains. (A) Multiple sequence alignment that overlays secondary structure regions. Homeodomain like proteins in our dataset: EnHD (1enh), hTRF1 (1ba5), c-Myb (1idy) and hRAP1(1fex); peripheral subunit binding domains: BBL (2bth), E3BD (1w4e) and POB (1w4j). (B) Optimal superposition of EnHD and BBL (left and right respectively, both in cartoon representation) to the members of their superfamilies (ribbon). Coloring as in A. (C) Comparison of values of local stabilization energy with predictions of AGADIR.

AGADIR also predicts 10-fold differences in helix propensity, which once again are correlated with the differences in local energy obtained from our analysis (Fig. 3C).

The correlation between the local energies from our analysis and the AGADIR helix propensities strongly supports the validity of calculating folding–unfolding kinetics from the one-dimensional free energy surfaces generated by the mean-field model. This result further buttresses the idea that folding speed is ultimately determined by the contribution from local interactions to the protein stabilization energy. For proteins of the same size and 3D structure the free energy barrier to folding is thus inversely proportional to the contribution from local interactions to their stabilization energy, which in turn is determined by relatively small differences in amino acid sequence. Because the total stabilization energy is nearly constant, it follows that faster folding proteins have smaller non-local energy terms.

Local and non-local energies *versus* three-dimensional structures

The exact fits to the model offer an invaluable opportunity to investigate the connection between stabilization energy and native 3D structure. The role of native structure in protein folding is typically modeled using a native-centric potential. There are many different algorithms to derive native interactions from 3D structures, varying among other things in the definition of the contacting atoms, the distance cut-off, and the counting rules. But, in virtually all native-centric modeling efforts the energy is assumed to be proportional to the number of residue–residue structural contacts, however they might be calculated.[3,5,20,21,25,35,42,57] Here we test this assumption comparing the number of native contacts observed in protein 3D structures with the energies extracted from the model analysis. Our focus is a general comparison rather than an exhaustive analysis of different contact descriptions. We thus count contacts following the most standard description that is employed in parameterizing Go models:[9] we define a native contact between any pair of residues when there is at least one atom pair at a distance closer than 5 Å. We then classify the observed contacts as local when the residues are separated by a maximum of three residues in sequence (i, i + 3), and non-local for all the others.

Fig. 4A shows the per-residue comparison between the empirical energies and the number of local and non-local contacts for the 52 proteins of our analysis. As before, the data are colored according to the three basic structural classes. The first observation is that there is no apparent correlation between the two properties when compared at the individual level, as indicated by the fact that the scatter of data points is as large as the full dynamic range of the correlation for both local and non-local contacts. This somewhat surprising result suggests that after removing size effects there is no much more information about the folding energetics of individual proteins that can be extracted from the structural variability in native contacts.

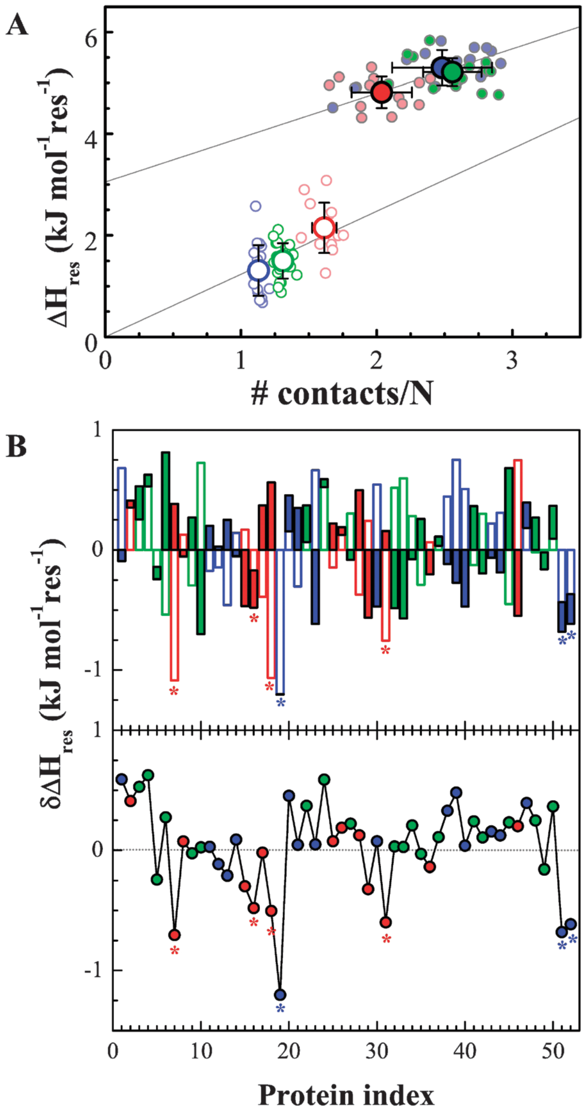

Fig. 4 Relationship between the native contacts and the stabilization energy. (A) Average values of the stabilization free energy per residue *vs.* the density of native contacts for the local (empty circles) and non-local (full circles) for the different structural types. Error bars represent one standard deviation on the mean value. Linear fits for local and non-local contributions are shown in grey. We show in blurred colors the values for the individual proteins. (B) Top: Difference between fitted and predicted local and non local stabilization energies per residue using the number of native contacts. Marked with asterisks, left to right: BBL, E3BD, EnHD, FBP28, POB, WW prototype and Yap. Bottom: Compensation between excess local and non-local stabilization energies.

There are some trends at the structural class level. The trends become more apparent if we calculate the average values within each of the three structural classes (Table 3). This confirms that α-helical proteins have higher local stabilization energies and also more local contacts. β-sheet proteins are the opposite, exhibiting large non-local stabilization energies and more non-local contacts. α + β proteins are somewhere in between. In fact, despite the relatively small differences between structural classes, the class-averaged properties correlate for both the local and non-local cases (see Fig. 4A). The correlation line between local energies and contacts crosses the origin, indicating that there is a direct proportionality between the number of contacts and the energy. In this case the scatter occurs mostly in the vertical direction, which implies that among proteins of a given class there is little structural variability but large fluctuations in local energy. Thus the differences in local energies within classes are likely due to sequence selection. The scenario for non-local interactions is quite different. The correlation line shows an offset of ∼3 kJ mol^{-1} per residue of non-local energy that is not associated to contacts. This large contact-independent energy (amounts to more than 50% of the non-local contribution) could reflect the overall contribution to protein stability arising from the burial of hydrophobic surface upon formation of globular structures (hydrophobic effect). This is an important finding because the native-centric models used in coarse-grained simulations usually calculate the energy of the non-local term as a simple sum of the pair-wise interactions defined by the matrix of structural contacts. It is also noteworthy that the non-local energies show little variability within each structural class, whereas the numbers of non-local contacts fluctuate much more. That is, for non-local interactions the scatter is mostly along the horizontal axis, suggesting that non-local energies do not depend that much on the specific details of protein structures.

Table 3 Average values for the stabilization energy per residue and the number of contacts per residue for the proteins in our database

	ΔH_{res}/kJ mol^{-1} res^{-1}		# Contacts/*N*	
	Local	Non local	Local	Non local
α	2.15	4.82	1.61	2.04
β	1.31	5.30	1.13	2.48
α + β	1.50	5.21	1.31	2.56

In principle, the correlations shown in Fig. 4A can be used as conversion rules to parameterize the stabilization energies of the model for any given protein from the atomic coordinates of its 3D structure. However, the calculations we performed using such structure-based protein-specific energies do not improve the prediction of folding rates relative to the simplest calculation that uses a single ΔH_{res} and a fixed ratio of local and non-local interactions for all proteins.[28] Furthermore, the calculations with the protein-specific parameters predict large variations in protein stability, in clear disagreement with the empirical observations. The reason is that structural and energetic fluctuations are for the most part uncorrelated. The lack of correlation is apparent in Fig. 4B, which shows the difference between the stabilization energies predicted by the number of contacts and the values obtained from the exact fits separated in the local and non-local terms (upper panel) or summed up (lower panel). A positive value of $\delta\Delta H_{res}$ means that the protein has an excess of native contacts whereas a negative value means too few native contacts. The difference (in absolute value) between the structural prediction and the empirical values is on average ∼0.37 kJ mol^{-1} for the local contribution and ∼0.28 kJ mol^{-1} for the non-local energy. This corresponds to only about 5% of the mean total energy per residue (Fig. 2A). However, the effect in the prediction of folding and unfolding rates is large because any mismatch in the large energy–entropy compensation strongly affects the free energy barrier height. On the other hand, the mean difference in total energy (the sum of the two terms) is only ~0.28 kJ mol^{-1}, indicating that there is some compensation between the deviations in the local and non-local terms estimated from the number of contacts. That is, protein structures with too many contacts of one type (whether local or non-local) tend to compensate such excess with fewer contacts of the other type.

There are, however, a few notable exceptions. Particularly, the ultrafast folders from the peripheral subunit binding domain family (BBL, POB, E3BD), engrailed homeodomain and WW domains (FBP28, WW prototype and Yap) have much larger local stabilization energies than is predicted from their structures, whereas their non-local energies follow the general trend (marked with asterisks in Fig. 4B). These results indicate that ultrafast folding does not originate from capillarity effects but rather from strong sequence selection towards local stabilization energies that are much higher than expected from the 3D structures. Overall, the results summarized in Fig. 4 imply that, while there clearly is a connection between stabilization energy and number of contacts per residue, the uncorrelated fluctuations of the two properties impede deriving accurate energies for individual proteins from their 3D structures.

Prediction of absolute folding and unfolding rates from size and structural class

The results discussed in the previous section suggest that introducing increasingly sophisticated descriptions of protein 3D structures is not a good strategy for improving rate predictions. However, the trends at the structural class level indicate that we can develop a simple physical model that captures the scaling effects of size and topology in both folding and unfolding rates. We introduced protein-class-sensitive energetics into the mean-field model by just defining the local ($\Delta H_{\text{loc,res}}$) and non-local ($\Delta H_{\text{non-loc,res}}$) contributions to the stabilization energy using the class-specific values of Table 3. We then implemented this version of the model into the computer algorithm PREFUR (PREdiction of Folding and Unfolding Rates), which calculates the 1D free energy surface at room temperature of any given single-domain protein from just its size (N) and ascription to one of the three structural classes (α, β, α + β). This is equivalent to using just 10-bits of protein specific information (N up to 256 residues plus three structural classes). The calculated folding free energy surface is then used by PREFUR to predict the absolute folding and unfolding rates from the height of the free energy barrier in both directions. The algorithm also estimates the confidence range in the predictions by calculating free energy surfaces obtained with local and non-local contributions that are one standard deviation away from the mean values shown in Table 3. This confidence range represents the estimated rate variability due to the sequence-dependent fluctuations in the local and non-local energy contributions.

To test the predictive performance of PREFUR we used again the database of 52 proteins and a cross-validation scheme in which we iteratively calculated the local and non-local energies for each protein by excluding it from the database. The prediction of absolute folding rates is shown in Fig. 5A. This figure demonstrates that PREFUR calculates absolute folding rates that strongly correlate with the experimental values (linear correlation coefficient in logarithmic scale of R = 0.78) with a near 1 to 1 slope. This correlation coefficient is slightly better than what one of us obtained before for size alone on a database with much larger spread in protein size (from 15 residue peptides to over 300 residue proteins).[18] The narrower range of protein sizes in this database constitutes a much more demanding test for the model. From bootstrap analysis we estimate a confidence range of 0.70–0.86 for the correlation coefficient. In practical terms this result implies that absolute folding rates can be predicted with an uncertainty of ±0.7 log units. Comparison with the predictions of a number of structural descriptors[3,20–22,24,42,43] shows that PREFUR is as good as the best performing one, which for this database is the long range order (LRO) (see Table 2). In this case matching the best structural predictor implies a significant improvement because PREFUR requires much less protein specific information (10-bits *versus* the atomic resolution 3D structure) and successfully predicts the absolute range in experimental folding rates. Moreover, the absolute unfolding rates predicted by

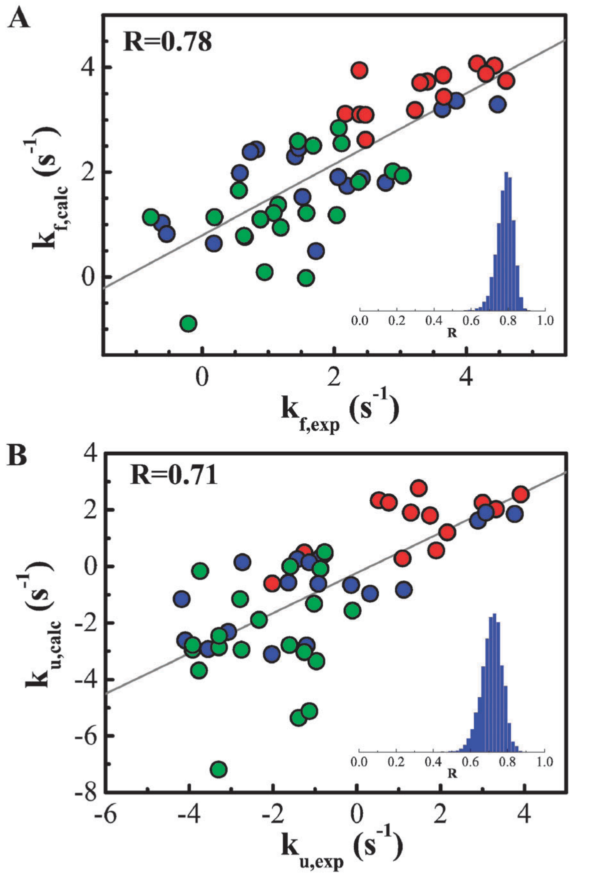

Fig. 5 Prediction of folding and unfolding rates from size and structural type. Prediction of protein folding (A) and unfolding (B) rates obtained using the average values of $\Delta H_{\text{res,loc}}$ and $\Delta H_{\text{res,non–loc}}$ for each structural type shown in Table 3. Insets are histograms of correlation coefficients from jack-knife tests. Correlation lines are shown in gray.

PREFUR also show a very good correlation with the experimental values, rendering a correlation coefficient of $R = 0.71$ in logarithmic scale (Fig. 5B), a confidence range of 0.59–0.82, and an uncertainty of ±1.4 log units. This is a remarkable result because unfolding rates have been more difficult to predict than folding rates. A closer inspection of Fig. 5 reveals that there are three $\alpha + \beta$ proteins for which the prediction performance is particularly poor, with both folding and unfolding rates being grossly under-predicted (removing these proteins from the analysis increases the correlation coefficients to 0.83 for folding and 0.80 for unfolding). Coincidentally, these three proteins are by far the three largest domains in the database (Azurin with 128 residues, CheW with 151, and Tm1023 with 125), suggesting that for such large sizes folding becomes organized in subdomains so that the mean-field approximation starts to break down leading to an overestimation of the free energy barrier in both directions. Finally, it is important to note that, because PREFUR performs an integrated calculation of folding and unfolding rates from the 1D free energy surface, it does also predict the experimental protein stabilities with an uncertainty of ±6.3 kJ mol^{-1}, which for this database is equivalent to less than 0.1 kJ mol^{-1} per residue.

Conclusion

In previous work we have used a one-dimensional free energy surface model to interpret kinetic experiments in ultrafast folding proteins,[29] rationalize the differences in equilibrium and kinetic behavior between structural homologues,[38] estimate thermodynamic folding barriers from calorimetric data,[58] and to investigate size-scaling effects in protein folding rates and stability.[28] The latter study indicated that all the experimental variability could be accounted for by fluctuations of only 1.7% and 8% in the two protein-specific parameters of the model (total stabilization energy per residue, ΔH_{res}, and curvature, $\kappa_{\Delta H}$).[28] This striking result has been recently interpreted as evidence that evolution operates with a very limited range of possibilities, even though protein folding rates vary by 6–7 orders of magnitude.[59] Here we have modified the mean-field model to explicitly account for the differences between local and non-local interactions and thus rationalize our previous result in structural terms. In the new implementation of the model the relative local and non-local contributions determine the shape of the free energy surface, including the height of the folding barriers, and, ultimately, the folding and unfolding rates (*i.e.* kinetics treated as diffusion on the free energy surface). By defining the interplay between local and non-local interactions as a major determinant of folding rates we use the same physical rationale used before by others.[22,32,34,35,60]

From the analysis of protein folding and unfolding rates with the modified model we extract individual contributions from local and non-local interactions for the 52 proteins included in the database, and observe that they vary quite significantly. However, such variability is strongly anti-correlated ($R = -0.97$), resulting in nearly constant total stabilization energies per residue. The anti-correlation between the local and non-local contributions to protein stability explains why the folding and unfolding rates change much more than protein stability.[28] In fact, the almost perfect compensation between these two largely varying terms indicates that there must be very strong evolutionary pressure towards a narrow stability range in globular proteins. According to this concept, the stability should be enough to guarantee folding, but not excessive to facilitate degradation after the end of the protein's duty cycle. However, individual proteins seem to attain this narrow stability range by different mechanisms depending on protein size, structural class, and sequence. Moreover, the large variability in the local/non-local ratio we observe for proteins of similar size and structure demonstrates that sequence selection provides a large tuning range for the local and non-local stabilization energies. Invoking maximum parsimony we can argue that the fact that some natural proteins have much higher local/non-local ratios than their structural homologues (but nearly the same total energy) is evidence of strong evolutionary pressure towards achieving downhill (*i.e.* barrier-less) folding in these proteins. Such a strong selection towards downhill folding supports the hypothesis of a biological role for the complex conformational behavior associated to this folding scenario.[61]

The comparison between the empirical local and non-local energies and the number of native contacts observed in protein 3D structures leads to important insights into the structure–energy relationships in protein folding. We show that the numbers of local and non-local native contacts are correlated to the empirical energies when they are compared at the structural-class level. This correspondence is summarized in the class-dependent structure–energy relationships of Table 3, which can be directly used for calibration of native-centric potentials and simulations with Go-like models. On the other hand, the variability in the number of contacts for proteins within each structural class is uncorrelated to the energies. Particularly, about half of the non-local energy per residue is independent of the number of non-local contacts (possibly reflecting the hydrophobic effect). The remainder non-local energy varies much less within members of one class than the number of contacts does. For local interactions there is large energetic variability within each class with almost no differences in the number of contacts. Accordingly, the best strategy for implementing native-centric potentials seems to be: (1) rescale the number of contacts observed in a given protein to energies following the class-dependent parameters per residue of Table 3, and, (2) introduce protein-specific local energies estimated from the amino-acid sequence (*e.g.* from the scores of secondary structure prediction algorithms).

Finally, by introducing class-dependent local and non-local stabilization energies into the mean-field model we show that is possible to predict absolute folding and unfolding rates in an integrated fashion using protein size and structural-class ascription as only input (*i.e.* 10 bits). The prediction of folding rates is as accurate as that of the best performing structural descriptors (*LRO*). But, more importantly, the unfolding rates, which have traditionally been considered extremely difficult to predict, are calculated from the protein size and structural class with significant accuracy. Because the folding and unfolding rates are calculated in an integrated fashion from the free energy surface, our model also predicts protein stability at room temperature with an accuracy of ±6.3 kJ mol^{-1}. Our integrated prediction of absolute folding and unfolding rates builds on the results recently shown by others[27] and provides a critical milestone in the development of a quantitative theory for protein folding.

Obviously, there are many additional factors that affect protein energetics and are not considered in our overly simplistic analysis. However, our calculations demonstrate that the major factors controlling the folding landscapes of well-evolved globular proteins are size and structural class. From the analysis of the 52 proteins in our database we estimate that all other factors contribute ±0.7 and ±1.4 log units to the natural variability in folding and unfolding rates, respectively. Interestingly, these estimates are almost exactly twice as large as the experimental variability in folding (±0.38) and unfolding (±0.75) rates induced by single point mutations (determined from more than 800 mutants on 24 proteins).[47] This comparison leads to two important conclusions. First, the ratio between the experimental variability in unfolding and folding rates induced by mutation reflects the recently discovered universal partitioning of the stabilization energy between folding (1/3) and unfolding (2/3) that results in a global φ-value of $\sim$0.30.[47] The fact that the unfolding/folding prediction uncertainties have exactly the same ratio of 2 confirms that the variability in experimental rates that is not captured by PREFUR is due to specific energetic factors modulated by the sequence. Second, such sequence-dependent variability in rates is only two-fold more for proteins with vast differences in sequence, size and 3D structure than it is for single-point mutations. This result highlights again the presence of large compensations in folding energetics, and suggests that natural protein sequences have been selected by evolution to minimize their differences in folding behavior (mostly stability). Furthermore, the little customization of folding properties by sequence selection that is present in natural proteins seems to mostly be achieved through engineering of local interactions within each structural class. PREFUR thus emerges as a powerful bioinformatic tool for both, the integrated prediction of folding and unfolding rates, as well as for the analysis of the specific energetics of a given protein by fitting the model to the available experimental data. Both applications have been implemented into a web server that is openly available to the scientific community at the url: http://tmg.cib.csic.es/servers/PREFUR.

Acknowledgements

We thank Ravishankar Ramanathan for the implementation of the PREFUR web application. This work was supported by the Marie Curie Excellence Award MEXT-CT-2006-042334, and the grants BFU2008-03237 and CONSOLIDER CSD2009-00088 from the Spanish Ministry of Science and Innovation (MICINN). David De Sancho acknowledges a FEBS long-term postdoctoral fellowship.

Notes and references

1 J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Proteins: Struct., Funct., Bioinf.*, 1995, **21**, 167–195.
2 J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten and P. G. Wolynes, *Fold. Des.*, 1996, **1**, 441–450.
3 K. W. Plaxco, K. T. Simons and D. Baker, *J. Mol. Biol.*, 1998, **277**, 985–994.
4 H. Taketomi, Y. Ueda and N. Go, *Int. J. Pept. Protein Res.*, 1975, **7**, 445–459.
5 V. Muñoz and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 11311–11316.
6 C. Clementi, *Curr. Opin. Struct. Biol.*, 2008, **18**, 10–15.
7 R. D. Hills Jr. and C. L. Brooks 3rd, *Int. J. Mol. Sci.*, 2009, **10**, 889–905.
8 A. Zarrine-Afsar, S. Wallin, A. M. Neculai, P. Neudecker, P. L. Howell, A. R. Davidson and H. S. Chan, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 9999–10004.
9 J. Karanicolas and C. L. Brooks 3rd, *J. Mol. Biol.*, 2003, **334**, 309–325.
10 A. Badasyan, Z. Liu and H. S. Chan, *J. Mol. Biol.*, 2008, **384**, 512–530.
11 E. Alm and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 11305–11310.
12 O. V. Galzitskaya and A. V. Finkelstein, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 11299–11304.
13 M. S. Li, D. K. Klimov and D. Thirumalai, *Polymer*, 2004, **45**, 573–579.
14 D. Thirumalai, *J. Phys. I*, 1995, **5**, 1457–1467.
15 A. V. Finkelstein and A. Badretdinov, *Fold. Des.*, 1997, **2**, 115–121.
16 P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 6170–6175.
17 A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, *Phys. Rev. Lett.*, 1996, **77**, 5433–5436.
18 A. N. Naganathan and V. Muñoz, *J. Am. Chem. Soc.*, 2005, **127**, 480–481.
19 O. V. Galzitskaya, S. O. Garbuzynskiy, D. N. Ivankov and A. V. Finkelstein, *Proteins: Struct., Funct., Bioinf.*, 2003, **51**, 162–166.
20 D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker and A. V. Finkelstein, *Protein Sci.*, 2003, **12**, 2057–2062.
21 M. M. Gromiha and S. Selvaraj, *J. Mol. Biol.*, 2001, **310**, 27–32.
22 L. Mirny and E. Shakhnovich, *Annu. Rev. Biophys. Biomol. Struct.*, 2001, **30**, 361–396.
23 A. Y. Istomin, D. J. Jacobs and D. R. Livesay, *Protein Sci.*, 2007, **16**, 2564–2569.
24 H. Zhou and Y. Zhou, *Biophys. J.*, 2002, **82**, 458–463.
25 K. W. Plaxco, K. T. Simons, I. Ruczinski and D. Baker, *Biochemistry*, 2000, **39**, 11177–11183.
26 M. M. Gromiha, S. Selvaraj and A. M. Thangakani, *J. Chem. Inf. Model.*, 2006, **46**, 1503–1508.
27 D. N. Ivankov and A. V. Finkelstein, *J. Phys. Chem. B*, 2010, **114**, 7930–7934.
28 D. De Sancho, U. Doshi and V. Muñoz, *J. Am. Chem. Soc.*, 2009, **131**, 2074–2075.
29 A. N. Naganathan, U. Doshi and V. Muñoz, *J. Am. Chem. Soc.*, 2007, **129**, 5673–5682.
30 N. D. Socci, J. N. Onuchic and P. G. Wolynes, *J. Chem. Phys.*, 1996, **104**, 5860–5868.
31 A. D. Robertson and K. P. Murphy, *Chem. Rev.*, 1997, **97**, 1251–1268.
32 V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, *J. Mol. Biol.*, 1995, **252**, 460–471.
33 V. Muñoz, E. R. Henry, J. Hofrichter and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 5872–5879.
34 V. Muñoz and L. Serrano, *Fold. Des.* , 1996, **1**, R71–77.
35 M. M. Gromiha and S. Selvaraj, *Biophys. Chem.*, 1999, **77**, 49–68.
36 A. Fung, P. Li, R. Godoy-Ruiz, J. M. Sanchez-Ruiz and V. Muñoz, *J. Am. Chem. Soc.*, 2008, **130**, 7489–7495.
37 P. Li, F. Y. Oliva, A. N. Naganathan and V. Muñoz, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 103–108.
38 A. N. Naganathan, P. Li, R. Perez-Jimenez, J. M. Sanchez-Ruiz and V. Muñoz, *J. Am. Chem. Soc.*, 2010, **132**, 11183–11190.
39 S. J. DeCamp, A. N. Naganathan, S. A. Waldauer, O. Bakajin and L. J. Lapidus, *Biophys. J.*, 2009, **97**, 1772–1777.
40 K. L. Maxwell, D. Wildes, A. Zarrine-Afsar, M. A. De Los Rios, A. G. Brown, C. T. Friel, L. Hedberg, J. C. Horng, D. Bona, E. J. Miller, A. Vallee-Belisle, E. R. Main, F. Bemporad, L. Qiu, K. Teilum, N. D. Vu, A. M. Edwards, I. Ruczinski, F. M. Poulsen, B. B. Kragelund, S. W. Michnick, F. Chiti, Y. Bai, S. J. Hagen, L. Serrano, M. Oliveberg, D. P. Raleigh, P. Wittung-Stafshede,

S. E. Radford, S. E. Jackson, T. R. Sosnick, S. Marqusee, A. R. Davidson and K. W. Plaxco, *Protein Sci.*, 2005, **14**, 602–616.

41 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

42 D. E. Makarov and K. W. Plaxco, *Protein Sci.*, 2003, **12**, 17–26.

43 D. N. Ivankov and A. V. Finkelstein, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 8942–8944.

44 H. Kaya and H. S. Chan, *Proteins: Struct., Funct., Bioinf.*, 2003, **52**, 524–533.

45 E. Krissinel and K. Henrick, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2004, **60**, 2256–2268.

46 H. S. Chan and K. A. Dill, *Proteins: Struct., Funct., Bioinf.*, 1998, **30**, 2–33.

47 A. N. Naganathan and V. Muñoz, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 8611–8616.

48 U. Mayor, J. G. Grossmann, N. W. Foster, S. M. Freund and A. R. Fersht, *J. Mol. Biol.*, 2003, **333**, 977–991.

49 I. E. Sanchez and T. Kiefhaber, *J. Mol. Biol.*, 2003, **327**, 867–884.

50 M. Sadqi, D. Fushman and V. Muñoz, *Nature*, 2006, **442**, 317–321.

51 N. Ferguson, C. M. Johnson, M. Macias, H. Oschkinat and A. Fersht, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 13002–13007.

52 A. N. Naganathan, U. Doshi, A. Fung, M. Sadqi and V. Muñoz, *Biochemistry*, 2006, **45**, 8466–8475.

53 F. Liu and M. Gruebele, *J. Mol. Biol.*, 2007, **370**, 574–584.

54 R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **107**, 1088–1093.

55 V. Muñoz and L. Serrano, *Nat. Struct. Biol.*, 1994, **1**, 399–409.

56 S. Gianni, N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. White, M. L. DeMarco, V. Daggett and A. R. Fersht, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 13286–13291.

57 V. Tozzini, *Curr. Opin. Struct. Biol.*, 2005, **15**, 144–150.

58 A. N. Naganathan, J. M. Sanchez-Ruiz and V. Muñoz, *J. Am. Chem. Soc.*, 2005, **127**, 17970–17971.

59 J. J. Portman, *Curr. Opin. Struct. Biol.*, 2010, **20**, 11–15.

60 R. Doyle, K. Simons, H. Qian and D. Baker, *Proteins: Struct., Funct., Bioinf.*, 1997, **29**, 282–291.

61 M. M. Garcia-Mira, M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz and V. Muñoz, *Science*, 2002, **298**, 2191–2195.

62 N. Taddei, F. Chiti, P. Paoli, T. Fiaschi, M. Bucciantini, M. Stefani, C. M. Dobson and G. Ramponi, *Biochemistry*, 1999, **38**, 2135–2142.

63 N. A. van Nuland, F. Chiti, N. Taddei, G. Raugei, G. Ramponi and C. M. Dobson, *J. Mol. Biol.*, 1998, **283**, 883–891.

64 V. Villegas, J. C. Martinez, F. X. Aviles and L. Serrano, *J. Mol. Biol.*, 1998, **283**, 1027–1036.

65 N. Ferguson, T. D. Sharpe, P. J. Schartau, S. Sato, M. D. Allen, C. M. Johnson, T. J. Rutherford and A. R. Fersht, *J. Mol. Biol.*, 2005, **353**, 427–446.

66 S. Sato, T. L. Religa and A. R. Fersht, *J. Mol. Biol.*, 2006, **360**, 850–864.

67 S. E. Jackson and A. R. Fersht, *Biochemistry*, 1991, **30**, 10428–10435.

68 H. M. Rodriguez, D. M. Vu and L. M. Gregoret, *Protein Sci.*, 2000, **9**, 1993–2000.

69 D. Perl, C. Welker, T. Schindler, K. Schroder, M. A. Marahiel, R. Jaenicke and F. X. Schmid, *Nat. Struct. Biol.*, 1998, **5**, 229–235.

70 T. Schindler, M. Herrler, M. A. Marahiel and F. X. Schmid, *Nat. Struct. Biol.*, 1995, **2**, 663–673.

71 R. Chu, W. Pei, J. Takei and Y. Bai, *Biochemistry*, 2002, **41**, 7998–8003.

72 K. W. Plaxco, C. Spitzfaden, I. D. Campbell and C. M. Dobson, *J. Mol. Biol.*, 1997, **270**, 763–770.

73 E. R. Main, K. F. Fulton and S. E. Jackson, *J. Mol. Biol.*, 1999, **291**, 429–444.

74 K. W. Plaxco, J. I. Guijarro, C. J. Morton, M. Pitkeathly, I. D. Campbell and C. M. Dobson, *Biochemistry*, 1998, **37**, 2529–2537.

75 N. A. Van Nuland, W. Meijberg, J. Warner, V. Forge, R. M. Scheek, G. T. Robillard and C. M. Dobson, *Biochemistry*, 1998, **37**, 622–637.

76 L. Hedberg and M. Oliveberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 7606–7611.

77 J. I. Guijarro, C. J. Morton, K. W. Plaxco, I. D. Campbell and C. M. Dobson, *J. Mol. Biol.*, 1998, **276**, 657–667.

78 E. L. McCallister, E. Alm and D. Baker, *Nat. Struct. Biol.*, 2000, **7**, 669–673.

79 D. E. Kim, C. Fisher and D. Baker, *J. Mol. Biol.*, 2000, **298**, 971–984.

80 J. C. Horng, J. H. Cho and D. P. Raleigh, *J. Mol. Biol.*, 2005, **345**, 163–173.

81 D. E. Otzen, O. Kristensen, M. Proctor and M. Oliveberg, *Biochemistry*, 1999, **38**, 6499–6511.

82 A. R. Viguera, J. C. Martinez, V. V. Filimonov, P. L. Mateo and L. Serrano, *Biochemistry*, 1994, **33**, 2142–2150.

83 V. P. Grantcharova and D. Baker, *Biochemistry*, 1997, **36**, 15685–15692.

84 R. Guerois and L. Serrano, *J. Mol. Biol.*, 2000, **304**, 967–982.

85 S. J. Hamill, A. E. Meekhof and J. Clarke, *Biochemistry*, 1998, **37**, 8071–8079.

86 N. Schonbrunner, K. P. Koller and T. Kiefhaber, *J. Mol. Biol.*, 1997, **268**, 526–538.

87 J. Clarke, E. Cota, S. B. Fowler and S. J. Hamill, *Structure*, 1999, **7**, 1145–1153.

88 M. Silow and M. Oliveberg, *Biochemistry*, 1997, **36**, 7633–7637.