Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model

2006 J. Phys.: Condens. Matter 18 S307

(http://iopscience.iop.org/0953-8984/18/14/S14)

J. Phys.: Condens. Matter **21** (2009) 279801 (1pp) doi:10.1088/0953-8984/21/7/279801

Corrigendum

Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model
S Wallin and H S Chan 2006 *J. Phys.: Condens. Matter* **218** S307-S328

In table 1 on page S313, for Coicilin E9 immunity protein (PDB id 1imq), the chain length N should be 86 and the folding rate k_f should be $1.5 \times 10^3 s^{-1}$. These errors were merely typographical and had no effect on the results and conclusions of the paper.

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **18** (2006) S307–S328 doi:10.1088/0953-8984/18/14/S14

Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model

Stefan Wallin and Hue Sun Chan[1]

Department of Biochemistry, and Department of Medical Genetics and Microbiology, Faculty of Medicine, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada

E-mail: wallin@arrhenius.med.toronto.edu and chan@arrhenius.med.toronto.edu

Received 5 October 2005, in final form 3 November 2005
Published 24 March 2006
Online at stacks.iop.org/JPhysCM/18/S307

Abstract

The 'topology' of a protein native structure refers to the pattern of non-covalent contacts among its amino acid residues. Diverse folding rates of natural small single-domain proteins are known to correlate well with simple parameters derived from these patterns. Here we extend our investigation of possible physical underpinning of this remarkable topology–rate relationship by applying continuum Gō-like C_α Langevin modelling to 13 small proteins. Folding rates simulated at transition mid-points are well correlated with $\Lambda_{\rm D}$, a 'topomer search model' (TSM) parameter which equals the number of nonlocal contacts in a protein's native structure. This modelling success in mimicking experimental topology–rate relationships is largely a conformational entropic effect: while transition states are results of large entropy–energy compensations, the trend of variation of the activation free energy $\Delta G^\ddagger$ versus $\Lambda_{\rm D}$ in the model is dominated by $\Delta G^\ddagger$'s entropic component. Interestingly, the activation conformational entropy $\Delta S^\ddagger$ is well correlated (negatively) with the Boltzmann-averaged number of nonlocal contacts $\Lambda_{\rm D}^\ddagger$ in the putative transition state ensemble. Thus, for the present Gō-like explicit-chain models, $\Lambda_{\rm D}$'s ability to predict rates is rooted in its correlation with $\Lambda_{\rm D}^\ddagger$. However, the model transition states are much more diffuse than that postulated by TSM because $\Lambda_{\rm D}^\ddagger$ is significantly smaller than $\Lambda_{\rm D}$.

1. Introduction

Important reasons abound for studying the kinetics of protein folding and unfolding. In conjunction with thermodynamic measurements, kinetic data provide crucial clues for understanding how the physico-chemical driving forces, molecular dynamics and solvation

[1] Author to whom any correspondence should be addressed.

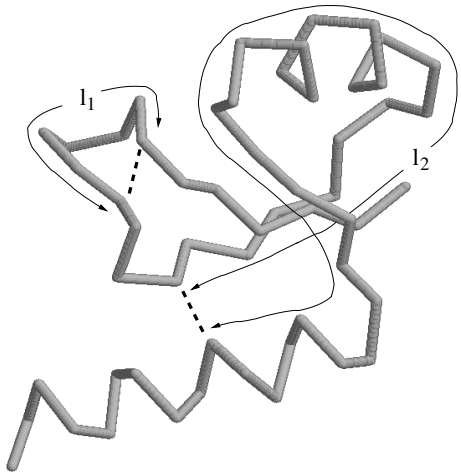

Figure 1. Definitions of contact orders and local and nonlocal contacts, using the native structure of the N-terminal domain of the protein L9 as an illustration. Amino acid residues are numbered consecutively along the chain sequence; contact order is defined to be $|i - j|$ for a contact between residues i and j. Among the many contacts in the structure shown in the figure, two are marked by dashed lines. A contact is local if it is made between amino acid residues close to each other along the chain sequence (i.e., low contact order, as for l_1). A contact is nonlocal if it is made between amino acid residues far apart along the chain sequence (i.e., high contact order, as for l_2).

effects lead to the emergence of a protein's native structure from its disordered unfolded conformations; and how disease-causing misfolding may occur. One intriguing puzzle confronting researchers of protein folding/unfolding kinetics has been the great diversity of folding rates observed for different proteins. In 1998, a key empirical advance that holds tremendous promise for shedding light on the physical origin of this phenomenon was made by Plaxco, Simons and Baker. They discovered that a simple parameter termed the relative contact order (CO), which is readily computable from the atomic coordinates of the native structure of a protein, can predict the folding rates of small single-domain 'two-state' proteins to within about one to two orders of magnitude [1]. The relative precision achieved by this empirical correlation is remarkable in view of the fact that the range of folding rates of such proteins varies over at least six orders of magnitude (from μs^{-1} to s^{-1}) [2]. By virtue of its fundamental nature, this finding of Plaxco *et al* presents an opportunity as well as a challenge to theoretical investigations [3], especially for 'big-picture' approaches that utilize simplified or so-called minimalist modelling [4–8].

1.1. Meanings of 'topology'

Plaxco *et al* use the term 'topology' for a protein's fold pattern as defined by its set of native contacts, i.e., pairs of amino acid residues that are in close spatial proximity in the native structure. Accordingly, the CO parameter is seen as a measure of 'topological complexity' [1]. Since the definition of topology is not uniform in the broad areas of biomolecular research covered in this special issue, a few words on the background of its present usage are in order. The terminologies of 'topological neighbour' and 'contact order' were introduced to the study of proteins in the late 1980s (figure 1). In that context, a pair of residues are called 'topological neighbours' to distinguish them from 'connected neighbours': topological neighbours are not covalently bonded along the chain sequence but are nonetheless spatially close to each

other in a given chain conformation [9]. The term topology is thus associated with non-covalent intraprotein contacts. This focus on contact patterns and the application of related ideas to lattice models at that time [10] have led to the hypothesis that compactness of chain conformation is a major driving force for the preponderance of secondary structure (helices and sheets) in proteins [11] as well as in crystals of synthetic polymers [12]. The viability of this early lattice model deduction is supported by more recent results from an elegant tube model formulation [13, 14],[2] although it has since been recognized that factors other than chain compactness alone—hydrogen bonding for example—have to be considered to account for the particular forms of secondary structure in real proteins [14–16]. Therefore, inasmuch as topology is used (as in the present work) to refer to contact patterns such as secondary structure in proteins, topology can be changed by conformational transitions without breaking any covalent bond. Consequently, while this meaning of topology is akin to that in the usage of the same term to describe RNA structural motifs such as pseudoknots [17], its meaning is significantly different from that of the topology of knots and links in DNA [18–20] or in the analyses of possible knotting and catenation in proteins [21–23].

1.2. Topological parameters in protein folding

Since the seminal work of Plaxco *et al* [1], several other topology-based parameters (topology as defined above by intraprotein native contacts) have been proposed. They have similar abilities to predict folding rates (k_f). These include long range order (LRO) [24], total contact distance [25], 'cliquishness' [26], local secondary structure content [27] and the topomer search model parameter Λ_D [28]. All of these parameters show good empirical correlation with the logarithm of the experimental folding rates ($\ln k_f^{exp}$). More recently, instead of using known native structures as starting points, direct predictions of k_f from sequence information alone have also been attempted. This is achieved by combining the proven predictive power of some of the topological parameters with bioinformatic algorithms for predicting secondary structure [29] and nonlocal contacts (see figure 1) [30] from amino acid sequences.

Although the success of these parameters in rate prediction hints at a certain simplicity in protein folding processes [31], it has proven nontrivial to devise a plausible physical picture that rationalizes the observed topology–rate relationship through an explicit account of the protein's intrachain interactions and conformational freedom. This difficulty notwithstanding, several theoretical treatments that do not consider explicit-chain representations have had remarkable successes in producing topology–rate relationships that are quantitatively similar to that observed experimentally [32–35]. These results are encouraging; but their physical implications have yet to be better elucidated by incorporation of their key modelling assumptions in self-contained, explicit-chain polymer models [36, 37]. In this regard, the topomer search model (TSM) [28, 38–40], at least in its current form, also belongs to this class of non-explicit-chain constructs. Similar to earlier non-explicit-chain treatments, TSM produces a good correlation between its topological parameter and logarithmic folding rate, indicating that the essential physics of folding must have been captured by this approach. However, the physical picture offered so far by the *mechanistic interpretation* of TSM is problematic [41, 42]. In particular, an explicit-chain analysis has raised fundamental concerns about the TSM assertion that the rate-limiting step in the folding of small single-domain proteins is an essentially unbiased search for the native topomer state [41].

[2] Self-avoiding walks on simple cubic lattices automatically satisfy the constraints on local and nonlocal triplets required in the tube model [18, 109]. HSC is indebted to Professor Micheletti for making this insightful observation during their discussion in a CECAM workshop held in May 2002 in Lyon, France.

1.3. Explicit-chain modelling and transition states

For models that use explicit representations of the protein chain to directly simulate folding kinetics, the goal of producing a set of k_fs that adequately captures the experimental trend of topology–rate correlation has been more elusive [43–45]. Nonetheless, several recent findings are noteworthy [46–50]. These advances include recognizing the pivotal role of thermodynamic and kinetic cooperativity in the topology–rate correlation among two-state proteins [8, 46, 47], and the discovery that certain many-body interactions [47, 49] such as local–nonlocal coupling [8, 47] can significantly improve a model's topology–rate relationship toward better agreement with experimental behaviour. Furthermore, a recent study of common native-centric chain models has provided intriguing evidence that correlation between model and real folding rates may be stronger if both the simulated and experimental rates are determined at their respective transition mid-point temperatures [48]. Taken together, these results have reinforced our expectation that simplified chain models can go a long way toward deciphering the physical basis of topology-dependent protein folding.

A key to physical understanding of the topology–rate correlation is to elucidate the relationship between various topological parameters and the rate-limiting step, or the transition state, in the folding process [51–56]. As local native contacts tend to speed up folding [1, 3], it has long been recognized that conformational entropy must be a dominant factor in the emergence of topology-dependent folding rates [1]. Pursuing this reasoning, Bai *et al* have studied the relationship between the 'total contact distance' [25] topological parameter and the 'size' of the folding transition state (critical nucleus) in a recent non-explicit-chain treatment [57]. Also, the effect of native topology on the breadth of the conformational space sampled by a folding protein, as characterized by a 'route measure', has been explored in the explicit-chain investigation of Chavez *et al* [48]. However, the impact of native topology on the conformational entropy of the transition state has not been clearly delineated. The present work uses explicit-chain modelling to address this critical issue.

As a first step in this endeavour, we construct native-centric C_α models to simulate the folding of a set of small single-domain proteins. Simplified, coarse-grained modelling is used here because currently even the most extensive atomic simulations [58–60] are far from providing sufficient and computationally efficient conformational coverage to address many equilibrium and long timescale kinetic properties of interest. The goals of our investigation are the following: (i) determine the conformational properties of the folding transition states in these model proteins; (ii) dissect the free energy folding barrier into entropic and enthalpic components; (iii) explore the interplay of intrachain interactions and conformational entropy [61] and the role of entropy–energy compensation in setting the height of folding barriers; and (iv) evaluate the relationship between the entropic component of the folding free energy barrier and several native topological parameters including relative contact order (CO), long range order (LRO) and the TSM Λ_D.

2. Model and methods

The explicit-chain model in this investigation follows from that introduced several years ago by Clementi *et al* [62], a coarse-grained approach that has since been used for the study of many different proteins (see, e.g., [43, 63]). The model is based on a simplified version of the protein chain, with each amino acid residue represented by its C_α position. The model interaction scheme is native-centric, or Gō-like [64–66], in that the potential function is not based on physico-chemical principles that are general for all proteins but rather designs a different energy function for each different protein, so as to bias the chain conformations

towards the protein's known native structure. Somewhat surprisingly, despite the teleological nature of such constructs [67], they have offered many physical insights that otherwise would not have been straightforward to discern. (A discussion of the biophysical justifications and limitations of Gō-like models can be found on pages 912 and 913 of [63].)

The potential energy function of the present model is given by

$$E = \sum_{\text{bonds}} K_r(b_i - b_i^{\text{n}})^2 + \sum_{\text{angles}} K_\theta(\theta_i - \theta_i^{\text{n}})^2 + \sum_{\text{dihedrals}} \{K_\phi^{(1)}[1 - \cos(\phi_i - \phi_i^{\text{n}})] + K_\phi^{(3)}[1 - \cos 3(\phi_i - \phi_i^{\text{n}})]\} + \sum_{i<j-3}^{\text{native}} \epsilon\left[5\left(\frac{r_{ij}^{\text{n}}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}^{\text{n}}}{r_{ij}}\right)^{10}\right] + \sum_{i<j-3}^{\text{non-native}} \epsilon\left(\frac{r_{\text{rep}}}{r_{ij}}\right)^{12}, \quad (1)$$

where b_i, θ_i, ϕ_i and r_{ij} are, respectively, the virtual bond lengths, bond angles, torsion angles and C_α–C_α distance between residues i and j, and b_i^{n}, θ_i^{n}, ϕ_i^{n} and r_{ij}^{n} are the corresponding native values provided by the PDB structure of the given protein.

As before, the summation of the second-last term in the above equation is over contacts that belong to the native contact set of the given protein. Here we adopt the criterion that a pair of amino acid residues is part of the native contact set if and only if any two non-hydrogen atoms (these include atoms in the sidechains), one from each of the two amino acid residues, are within 4.5 Å in the PDB native structure. This definition is the same as that used by Chavez *et al* [48] (but is not exactly identical to that in several previous studies [43, 49, 62, 63, 66, 68–70]), and can apparently lead to well-behaved model behaviours such as bimodal free energy profiles. The native contact sets so defined are used to determine the topological parameters CO, LRO and Λ_{D} for the model proteins. As will be discussed further below, native contacts with contact orders lower than a certain nonlocality cut-off l_{c} are excluded from the computation of LRO and Λ_{D}. In this work, the contact order of each native contact appears only once in the summation expression for CO [1]. In other words, every native contact contributes equally in that they are not weighted by the number of atomic contacts between the two contacting residues (cf equations (1) and (2) of [71]). This convention is adopted here to allow for the generalization of the CO measure to non-native conformations in the present C_α model. In general, for any conformation, a native contact is taken to exist between residues i and j if the pair belongs to the native contact set and $r_{ij} < 1.2r_{ij}^{\text{n}}$ [41, 63]. We use $\epsilon = 1$ in the present study. The values of the parameters K_r, K_θ, $K_\phi^{(1)}$, $K_\phi^{(3)}$ and r_{rep} in equation (1) are identical to those in [63].

Kinetic properties of Gō-like protein models with potentials similar to that in equation (1) have been studied by Newtonian mechanics in the absence of solvent frictional effects [43, 62], whereby velocity rescaling [72] was used to maintain a constant simulation temperature. Here, following our previous approach, frictional forces are considered. To this end, simulation of folding kinetics and thermodynamic sampling are performed using Langevin dynamics [73], with time evolution governed by the equation

$$m\dot{v}(t) = F_{\text{conf}} - m\gamma v(t) + \eta(t), \quad (2)$$

where m, v, $\dot{v}$, F_{conf}, γ and η are, respectively, mass, velocity, acceleration, conformational force, friction (viscosity) and the random force. Units are chosen such that $m = 1$; F_{conf} is the negative gradient of the potential energy function in equation (1); and η is drawn from a Gaussian distribution, the variance of which is determined by the temperature of the system (see equation (3) in [63]). As in previous studies from our group [63], equation (2) is integrated using the velocity Verlet algorithm [73–75], with a time step $\delta t = 0.02$ and a friction coefficient $\gamma = 0.0125$. This choice of γ is motivated by the modelling requirement for

computational tractability, as this relatively small value of γ allows for efficient simulations of folding/unfolding kinetics. The resulting dynamics is underdamped, however[3]. It corresponds to the low friction case of Veitshans *et al* [73], with an effective solvent viscosity much lower than that expected of real water (see pages 20 and 21 of [73]). In using this model, our assumption is that, aside from a different timescale, the general trends obtained from the present Langevin set-up are applicable to situations with more realistic solvent effects. In view of present computational limitations, this approach is useful for making advances. Nonetheless, it should be recognized that many aspects of this assumption remain to be tested, as the relationship between results from explicit-solvent and low and high friction implicit-solvent simulations can be rather complex [76].

For real proteins, the effective solvent-mediated intrachain interactions are temperature dependent. Consequently, folding rates are generally non-Arrhenius [77–79] and the entropic component of the free energy barrier to folding contains not only conformational contributions but also entropic effects of the solvent-mediated interactions. It is possible to construct explicit-chain models for such experimental behaviours [80–82], including the phenomenon of enthalpic folding barriers which we have recently proposed to be a likely consequence of cooperative desolvation effects [83–86]. Here we choose to keep the model interaction potential in equation (1) temperature independent. As the focus of the present investigation is on the role of conformational entropy in topology-dependent folding, this choice serves to simplify the modelling logic as it ensures that all entropic contributions are attributable to conformational effects.

3. Apparent two-state proteins

The attention of the present effort is on small, single-domain proteins. Kinetic and equilibrium experiments have indicated that many such proteins fold in a 'two-state' manner [2]. This means that the folding of these proteins—when viewed macroscopically—may be seen as proceeding more or less directly from the unfolded (or denatured) state, D, to the native state, N, with minimal or essentially non-existent accumulation of intermediate conformational population during the process. About 30 such single-domain proteins have been identified [87] since chymotrypsin inhibitor 2 (2ci2) was first demonstrated in 1991 to exhibit apparent two-state behaviour [88]. We focus here on a set of 13 such proteins (table 1).

4. Experimental versus theoretical folding rates

We begin our analysis by ascertaining the extent to which the native-centric polymer model described above is able to reproduce experimental folding rates. In general, it has been noted that the behaviour of the present class of Gō-like models are more two-state-like near each model's folding/unfolding transition mid-point, T_{m} [48, 101]. These observations are consistent with previous findings from our group that although the folding behaviour of these models satisfies thermodynamic and kinetic two-state criteria at temperatures close to T_{m}, there are substantial deviations from kinetic two-state cooperativity at other temperatures as manifested by significant chevron rollovers [63, 102]. For these reasons, only folding rates simulated at the T_{m}s of the model proteins are considered in this study.

Figure 2 compares simulated and experimental folding rates (all experimental folding rates used in this paper are in units of s^{-1}). The folding statistics for each model protein is

[3] This may be illustrated by a test simulation of one of the model proteins studied here (see section 3). Using the present Langevin parameters, we find that the average values of $|m\gamma v|$ and $|F_{\mathrm{conf}} + \eta|$ for one of the C_α positions far from either chain end are 0.0099 and 9.1, respectively, indicating that damping is much weaker compared to the force exerted on the given position.

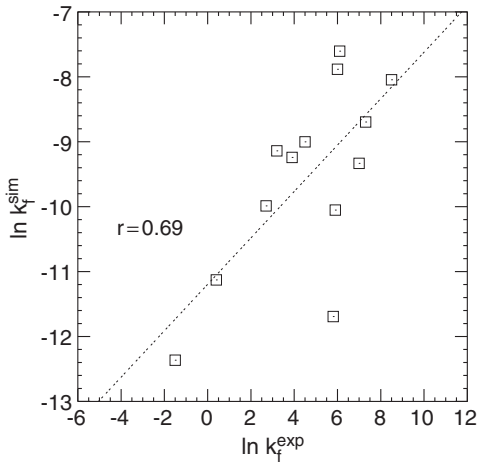

Figure 2. Folding rates for 13 proteins obtained from simulations of the present native-centric C_α model ($\ln k_f^{sim}$) are compared against the corresponding experimental folding rates measured at temperatures close to 25 °C in the absence of denaturant (data from table 1). In this and all subsequent scatter plots in this paper, r is the Pearson correlation coefficient, and straight lines are least-square linear fits to the data points.

Table 1. Data for the 13 proteins studied in this work. PDB id: Protein Data Bank accession code; N: number of amino acids; k_f: experimental folding rate in water (zero denaturant) determined at temperature T_{exp} [88–100]. The data for 1psf are from [87]; T_{exp} for 1psf is not available from this reference.

Protein	PDB id	N	k_f (s^{-1})	T_{exp} (°C)
Acylphosphatase	1aps	98	2.3×10^{-1}	28
Chymotrypsin inhibitor 2	2ci2	64	4.8×10^{1}	25
Spliceosomal protein U1A	1urn	96	3.2×10^{2}	25
λ-repressor	1lmb	80	4.9×10^{3}	25
SH3-domain (fyn)	1shf	59	9.0×10^{1}	20
Protein G	1pgb	56	4.0×10^{2}	22
Twitchin	1wit	93	1.5×10^{0}	20
CspB (*Bacillus subtilis*)	1csp	67	1.1×10^{3}	25
S6	1ris	97	3.7×10^{2}	25
Photosystem I accessory protein	1psf	69	2.5×10^{1}	—
N-terminal domain from L9	1div	56	4.5×10^{2}	19
Coicilin E9 immunity protein	1imq	59	9.0×10^{1}	10
His-containing phosphocarrier protein	1poh	85	1.5×10^{1}	20

obtained from 10^9 time steps of Langevin dynamics simulation after a short thermalization period. The number of unfolding/folding events recorded for the different proteins varies from 42 for the slowest-folding model protein to 4979 for the fastest-folding model protein. At the mid-point temperature T_m, the folding rate k_f is equal to the unfolding rate k_u, and kinetic relaxation is well approximated by a single exponential (figure 11(a) of [63]). This allows k_f to be calculated as the inverse of the mean first passage time (MFPT) from the denatured state to the native state, as well as vice versa, during the simulation; i.e., $k_f = (\text{MFPT})^{-1}$. Recent investigations using similar models and numbers of trajectories indicate that sampling errors for MFPTs obtained by this procedure are small [85, 86]. In the present calculation,

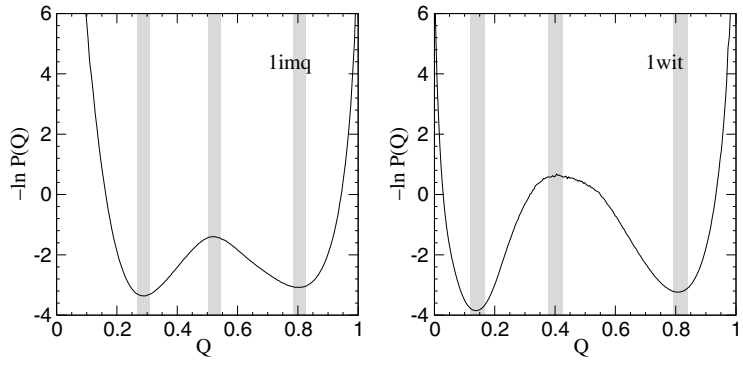

Figure 3. Free energy profiles $G(Q)/k_{\mathrm{B}}T_{\mathrm{m}} = -\ln P(Q)$, where k_{B} is the Boltzmann constant (set to unity in the present units), $P(Q)$ is the probability distribution of the fraction Q of native contacts formed. The profiles shown are for the proteins 1imq and 1wit (see table 1) at their respective model T_{m}s. The Q-width of each grey band is 0.05. They indicate the conformational spaces that we have used to define the denatured or unfolded (low Q), transition (intermediate Q) and native or folded (high Q) states.

we use the definitions of denatured (D, unfolded) and native (N, folded) states as defined in figure 3. Each of the free energy profiles in figure 3 shows a single barrier separating the denatured state (low Q minimum) and the native state (high Q minimum), indicating that the models behave roughly as two-state systems at their transition mid-point temperatures. All of the other 11 model proteins in our set exhibit similar bimodal free energy versus Q profiles. The match between the general trends for the simulated and experimental rates in figure 2 is reasonable (correlation coefficient = 0.69). However, the simulated rates span only approximately two orders of magnitude; thus they are much less diverse than the corresponding experimental rates that cover more than four orders of magnitude. Quite remarkably, the level of correlation obtained here between simulated and experimental folding rates is the same as that reported earlier by Koga and Takada [43]. They also computed folding rates at model T_{m}s (with simulated rates of 18 proteins spanning approximately 1.5 order of magnitude), but the set of proteins that they studied and their modelling set-up are not identical to ours.

At least two sets of ideas—which are not necessarily mutually exclusive—have been put forward to address this mismatch in folding rate diversity between experiments and explicit-chain model predictions. First, it has been pointed out that a probable culprit is the failure of many forms of Gō-like models to embody a sufficiently high degree of cooperativity [8, 46] to mimic real two-state proteins [63, 86]. Indeed, consistent with this general assessment, a local–nonlocal coupling lattice model interaction scheme that enhances cooperativity has been shown to also significantly increase the diversity in model folding rates [47]. But this local–nonlocal coupling principle has yet to be evaluated in off-lattice continuum contexts. Interestingly, a more recent study of a class of Gō-like chain models with three-body interactions as a 'perturbation' shows that such non-pairwise, many-body effects can significantly increase the range of model folding rates (some are deduced from thermodynamic considerations, some from direct kinetic simulations), and enhance the correlation between model and experimental rates [49]. Considered as a whole, these findings suggest strongly that cooperativity and non-additive many-body interactions are likely to be critical in accounting for the tremendous diversity among experimental folding rates.

Second, it should be noted that the experimental folding rates used in the present comparison are measured at temperatures close to $T = 25\,^{\circ}\mathrm{C}$ and at zero denaturant

concentration. Ideally, however, since the theoretical rates are simulated at model T_m (this choice is somewhat obligated by the limitations of common Gō models, as discussed above), it would appear more appropriate to compare these theoretical rates with real rates measured at the experimental T_m [48]. Unfortunately, currently those data are not available for a significant fraction of known two-state proteins, including some of the 13 proteins modelled here. In lieu of actual experimental T_m rate data for many of the proteins that they considered, Chavez *et al* [48] proposed a method for extracting a 'representative' folding rate of a protein from the folding rate data for its single-point mutants. For the set of two-state proteins that they considered, a remarkably high correlation ($r = 0.92$) between model and 'representative' rates was reported. But this method does not alleviate the problem of limited diversity among simulated rates. To address that, Chavez *et al* postulated an additional procedure for rescaling the simulated folding rates. The resulting rescaled rates were then shown to exhibit a diversity comparable to that of the experimental rates. However, while these new procedures appear promising, further analysis is required, as many pertinent issues remain to be better elucidated. Accordingly, we will comment briefly on their method for obtaining 'representative' rates in section 8.

5. Free energy barriers to folding: entropic and enthalpic components

We now direct our effort toward understanding the basis for topology–rate correlation in figure 2, and leave further elucidation of model rate diversity to future investigations. The correlation between experimental and simulated folding rates in figure 2 is not extremely high. Nonetheless, the very existence of a reasonable correlation—in conjunction with the established good correlation between native topology and experimental folding rate—implies that the present model is at least embodying part of the physics that underlies the experimentally observed topology–rate relationship. Hence, it is instructive to ask: what are the essential features of this explicit-chain model that allow it to capture this aspect of the folding of real proteins? To address this question, we focus on the properties of the rate-limiting folding transition state in the model. Understanding this state is important because for real proteins some of its properties are experimentally accessible through methods such as protein engineering and ϕ-value analysis [103].

Here we adopt a simple definition of the folding transition state. We define the transition state ensemble (TSE), similarly to how the D and N states are defined above for MFPT calculations, as the set of conformations having Q-values within a small region around the local maximum along the free energy profile $G(Q)$ between the D and N states [54, 55, 62], as illustrated in figure 3. A possibly more rigorous definition of TSE is via the quantity p_{fold}, the probability that a conformation reaches N before D, by identifying TSE with the set of conformations with $p_{\mathrm{fold}} = 1/2$ [52, 56, 104]. The present Q-based definition does not guarantee that all conformations in our TSE have $p_{\mathrm{fold}} = 1/2$. In that sense, our TSEs are putative, though the relationship between p_{fold} and rate-limiting events remains to be better elucidated. In this regard, it is clear that our TSE must be in large measure representative of the true rate-limiting step in the model, as it produces very good predictions of model folding rate at the transition mid-point. Figure 4 shows that the model folding rates are, up to a constant overall factor, almost completely determined by the height of the folding free energy barrier (i.e. activation free energy) $\Delta G^{\ddagger}$ computed using the values of $G^{\ddagger}$ and G^{D} which are readily read off from each of the 13 model proteins' free energy profile $G(Q)$ (cf figure 3). This is consistent with a lattice model observation that kinetic progress on a Q-based profile is quasi-continuous, and thus the thermodynamic Q-based free energy peak also amounts to a kinetic bottleneck (cf figure 2 of [55]). Similarly strong correlations between Q-based $\Delta G^{\ddagger}$

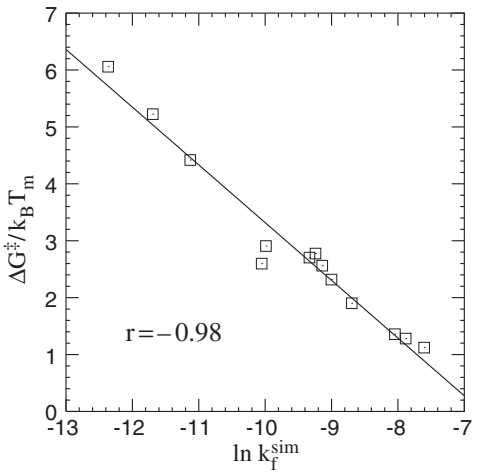

Figure 4. Correlation between simulated logarithmic folding rates, $\ln k_f^{sim}$, and the height of the free energy barrier to folding $\Delta G^{\ddagger} = G^{\ddagger} - G^{D}$ at the T_ms of the model proteins, where ‡ and D indicate, respectively, the transition state and denatured state.

and Gō-like model folding rate at T_m have also been reported before [48, 86]. Thus, these models' behaviour at T_m is apparently well described by Kramer's classic theory [105] which predicts a rate $\propto \exp(-\Delta G^{\ddagger}/k_B T)$. In other words, despite the obvious multi-dimensional character of the models' energy landscapes, a single one-dimensional parameter Q is able to capture a significant part of the rate-determining factors of the folding process.

To ascertain more specifically the role of conformational entropy in determining the folding rate, we dissect the free energy profile into its energetic and entropy components, namely,

$$G(Q) = \langle E \rangle_Q - T S(Q) \tag{3}$$

where $\langle E \rangle_Q$ is the average energy (referred to simply as the energy when the contextual meaning is clear) and $S(Q)$ is the entropy of the model protein, as functions of Q. The quantity $\langle E \rangle_Q$ is readily computed by collecting averages of the value taken by the potential energy function E in equation (1) from subsets of simulated conformations with different given Qs. Since $G(Q)$ is already known (up to an arbitrary additive constant), the $\langle E \rangle_Q$ computation also determines the entropy $S(Q)$. The activation energy $\Delta E^{\ddagger}$ and activation entropy $\Delta S^{\ddagger}$ can then be calculated as the differences, respectively, of the $\langle E \rangle_Q$ and $S(Q)$ values at the $\Delta G^{\ddagger}$-defined transition state minus that for the D state.

Figure 5 shows how $\Delta E^{\ddagger}$ and $\Delta S^{\ddagger}$ vary across the 13 model proteins. Not surprisingly, the results indicate large entropy–energy compensations. A large loss in conformational entropy (and thus a large *increase* in entropic free energy) is expected as folding proceeds (the value of the progress variable Q increases) because the chain conformations are becoming more compact. This is accompanied by a large decrease in energy because the compactifying chains are forming a larger number of favourable contacts. For instance, the slowest-folding model protein in figure 5 involves $\Delta E^{\ddagger}$ and $\Delta S^{\ddagger}$ of similarly large magnitudes of $\sim$70 $k_B T$; but the net $\Delta G^{\ddagger}$ value after compensation is only $\sim$6 $k_B T$. The logarithmic model folding rate $\ln k_f^{sim}$ correlates well with both $\Delta E^{\ddagger}$ and $\Delta S^{\ddagger}$, although these correlations are less strong than that between $\ln k_f^{sim}$ and $\Delta G^{\ddagger}$. Most interestingly, figure 5 shows that the trend of topology-dependent folding rate (which is governed by $\Delta G^{\ddagger}$; cf figures 2 and 4) is consistent with $\Delta S^{\ddagger}$ but opposite to that of $\Delta E^{\ddagger}$. In other words, the sign of the slope of variation of $\Delta G^{\ddagger}$

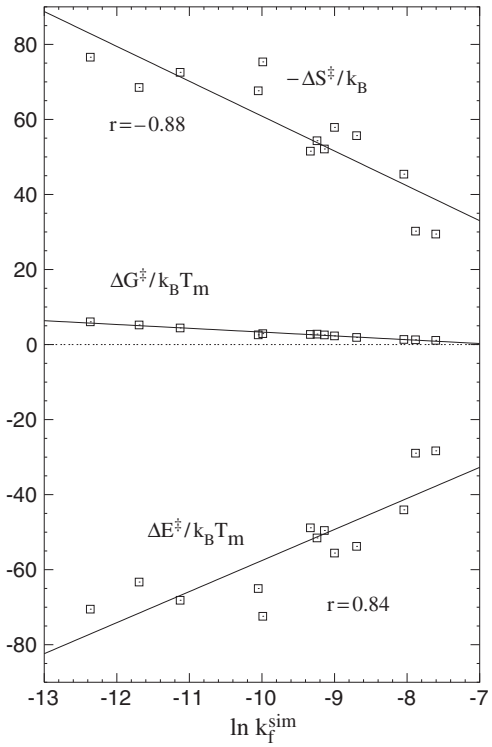

Figure 5. Activation free energy ($\Delta G^{\ddagger}$), activation energy ($\Delta E^{\ddagger} = \langle E\rangle^{\ddagger} - \langle E\rangle^{\mathrm{D}}$) and activation entropy ($\Delta S^{\ddagger} = S^{\ddagger} - S^{\mathrm{D}}$) at the T_{m}s of the 13 model proteins in this study. The strong correlation between $\ln k_{\mathrm{f}}^{\mathrm{sim}}$ and $-\Delta S^{\ddagger}$ ($r = -0.88$) and the even stronger correlation between $\ln k_{\mathrm{f}}^{\mathrm{sim}}$ and $\Delta G^{\ddagger}/T_{\mathrm{m}}$ ($r = -0.98$, cf figure 4) are significant because the probabilities that they have arisen from pure chance are very small. By scrambling the order of the 13 data points multiple times and recalculating the resulting correlation coefficients, we have determined that the probabilities of randomly achieving a correlation coefficient greater than or equal to the observed r values are less than 10^{-4} and 10^{-7}, respectively, for $\ln k_{\mathrm{f}}^{\mathrm{sim}}$ versus $-\Delta S^{\ddagger}$ and $\ln k_{\mathrm{f}}^{\mathrm{sim}}$ versus $\Delta G^{\ddagger}/T_{\mathrm{m}}$.

with respect to $\ln k_{\mathrm{f}}^{\mathrm{sim}}$ is identical to that of the entropic component, $-T\Delta S^{\ddagger}$, of $\Delta G^{\ddagger}$, but is opposite to that of the energetic component, $\Delta E^{\ddagger}$, of $\Delta G^{\ddagger}$. In this very sense, the topology–rate relationship, at least in the present model, is dominated by a conformational entropic effect. This suggests a folding process in which the search for the transition state is driven by favourable energetic interactions but the transition states of slow- and fast-folding proteins are associated, respectively, with low and high conformational entropies. Indeed, this conclusion is conceptually similar to the one advanced recently by Bai *et al* [57], although these authors did not use an explicit-chain approach in their investigation.

Figure 6 takes a closer look at the entropy–energy compensation phenomenon, using two model proteins as examples[4]. The upper panels of this figure show that large entropy–energy compensations are operative over the entire folding energy landscape, as the variations of $\langle E\rangle_Q/k_{\mathrm{B}}T$ and $S(Q)/k_{\mathrm{B}}$ over Q follow almost the same trend. The dependences of $\langle E\rangle_Q/k_{\mathrm{B}}T$ and $S(Q)/k_{\mathrm{B}}$ on Q are not linear; nonetheless, the general trends of their slope are very similar.

[4] Among the proteins considered here, 1imq is the second fastest experimentally, and fourth fastest in our model; whereas 1wit is second slowest experimentally, and third slowest in our model. Model free energy profiles of the faster-folding 1lmb and slower-folding 1aps have been presented elsewhere [41].

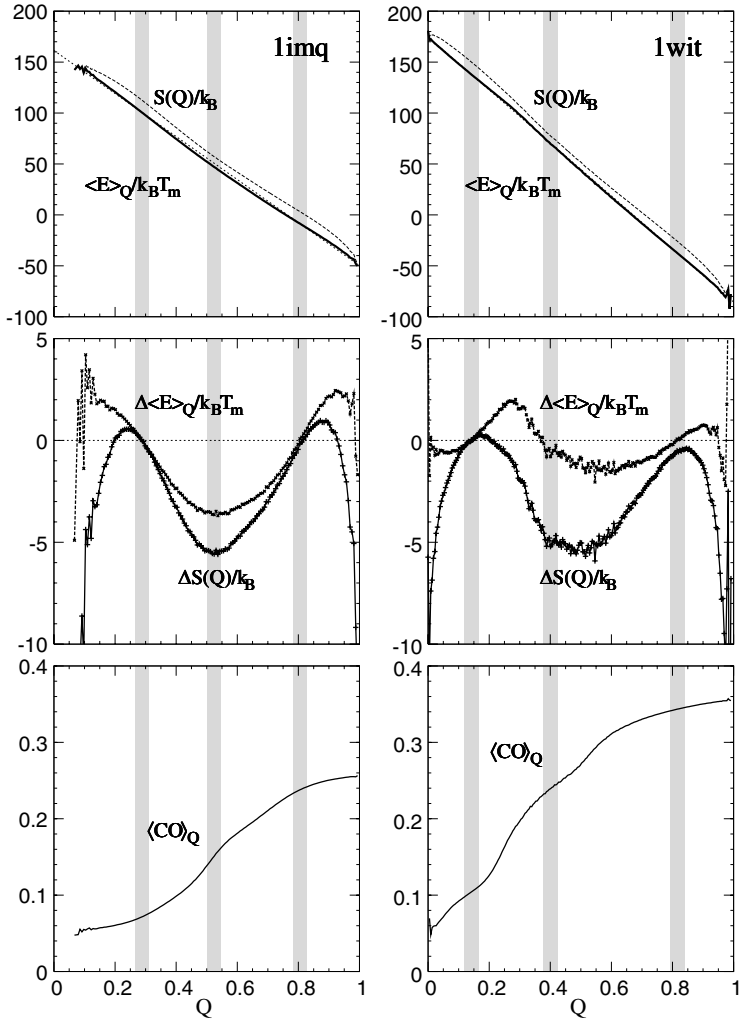

Figure 6. Entropy–energy compensation at $T_{\rm m}$ along the progress variable Q of a relatively fast-folding (1imq, left panels) and a relatively slow-folding (1wit, right panels) model protein. For each model protein, the shaded vertical bands mark (*from left to right*) the denatured, transition and native states, as in figure 3. Top panels: entropy ($S(Q)$, thick dotted curves) and energy ($\langle E\rangle_Q$, thick solid curves) as functions of Q, in the units shown. The thin dotted line (difficult to discern for 1wit) is a fitted straight-line approximation of the variation of $\langle E\rangle_Q$ with respect to Q. Middle panels: $\Delta\langle E\rangle_Q$ and $\Delta S(Q)$ are, respectively, the deviations of the actual simulated $S(Q)$ and $\langle E\rangle_Q$ values (as plotted as thick curves in the top panels) from the straight-line approximations (given by the thin dotted curves in the top panels). Subtractions of the energy and entropy curves in the upper, as well as the middle, panels result precisely in the two free energy profiles in figure 3. Bottom panels: the variation of the average value of relative contact order CO with respect to Q; $\langle{\rm CO}\rangle_Q$ is calculated by averaging over conformations with the given Q.

However, this entropy–energy compensation is not perfect. If the compensation were perfect, $G(Q)$ would be a constant and there would not be a free energy barrier to folding. The middle panels of figure 6 quantify this imperfection. The results indicate that entropy–energy compensation is less effective for the slower-folding model protein. More specifically, for the slower-folding protein with higher native state topological complexity (as characterized

by a larger Λ_D, for example), the decrease in energy that accompanies the gain in favourable interaction at the transition state is less capable of compensating for the concomitant loss in conformational entropy than that for the faster-folding protein. This is evident from the larger separation between the $\Delta\langle E\rangle_Q/k_BT_m$ and $\Delta S(Q)/k_B$ curves around the transition region at intermediate Q for 1wit than for 1imq in the middle panels of figure 6.

6. Characterizing transition states with simple topological parameters

To gain further insight into possible origins of topology-dependent folding, we extend the analysis of topological properties to all conformations and all Q values, as discussed in section 2, instead of restricting such considerations to just the $Q = 1$ native structure. In this way, Λ_D, LRO and CO may be treated as observables in the model. These parameters are all constructed to capture, in somewhat different ways, the 'locality' of the contacts in the native structure (figure 1). The relative contact order CO [1] is the average sequence separation of the native contacts divided by the total number of amino acid residues in the protein,

$$\mathrm{CO} = \frac{1}{NM}\sum_{ij} l_{ij} \tag{4}$$

where the sum is over all native contacts ij. Here N is the chain length, M is the total number of native contacts and the sequence separation l_{ij} of contact ij is the number of amino acid residues along the chain sequence between residues i and j (see section 2 and figure 1). The correlation between native CO values and the simulated folding rates of the 13 proteins studied here is reasonably good ($r = -0.59$; detailed data not shown), though it is somewhat lower than the correlation between CO and the experimental folding rate, e.g., the $r = -0.75$ reported by Ivankov *et al* for a larger set of proteins [87].

The bottom panels of figure 6 show the variation of CO along the Q-based free energy profile. They show a monotonic increase in $\langle\mathrm{CO}\rangle_Q$ with Q. This trend is expected because chain compactness and the number of intrachain contacts tend to increase with Q. For these examples, the dependence on Q of CO is sigmoidal for the faster-folding 1imq but not for the slower-folding 1wit. Further exploration of this feature is beyond the scope of the present study, but it would be interesting to study its implication in future work.

The Λ_D parameter [28] is the number of sequence-distant (nonlocal) contacts, where a contact ij is defined as nonlocal if residues i and j are separated by at least l_c residues, i.e., $|i - j| > l_c$. The definition of long range order LRO, which was introduced [24] before Λ_D, differs from Λ_D only by a normalization factor, i.e., $\mathrm{LRO} = \Lambda_D/N$, if both LRO and Λ_D are evaluated with the same sequence cut-off parameter l_c. Values from 4 to 12 have been suggested for this parameter [28]. We use $l_c = 12$ because a maximum correlation between LRO and the folding rate was observed using this cut-off in the original study of Selvaraj and Gromiha [24], and the main correlation result of Makarov and Plaxco was also obtained using $l_c = 12$ [28].

Figure 7 shows the dependences of our explicit-model folding rates on LRO and Λ_D. The correlation between $\ln k_f^{sim}$ and these two topological parameters is relatively good, with model correlation coefficients r (see figure 7) similar to that obtained from experimental folding rates (using different protein sets) for LRO ($r = -0.78$) [24] and for Λ_D ($r = -0.88$) [28][5]. Quite remarkably, the correlation between $\ln k_f^{sim}$ and Λ_D is very strong (figure 7(c)), suggesting that the present model system may be relatively well suited for exploring Λ_D dependence, even

[5] The correlation considered in [28] is between $\log(k_f/\Lambda_D)$ and Λ_D. The degree of correlation so obtained is essentially equal to that between $\ln k_f^{sim}$ and Λ_D because the negative correlation between Λ_D and $\ln\Lambda_D + \Lambda_D \ln a$ for $a = 0.86$ [28, 41] is almost perfect, with $r = -0.999\,890$, for example, for the set of 13 proteins in this study.

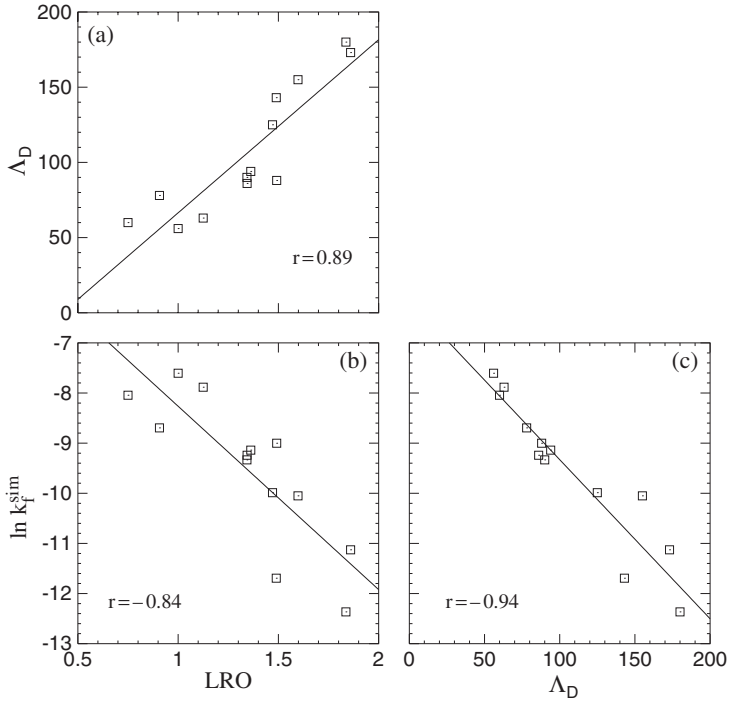

Figure 7. Correlation relationships between the folding rate k_f^{sim} of the present native-centric C_α model simulated at T_m and the topological parameters LRO (b) and Λ_D (c) computed for the native structures of the 13 proteins in this study. (a) shows the correlation between LRO and Λ_D for the same set of proteins.

though the correlation between $\ln k_f^{sim}$ and the logarithmic experimental folding rate $\ln k^{exp}$ is less strong (figure 2). Figure 7(a) highlights the high degree of similarity between LRO and Λ_D that follows from the definitions of these parameters (see above).

Figures 8–10 explores the relationship between native topology and the topological properties of the transition state, as well as the role of the latter in determining folding rates. In particular, we seek to better understand how native and transition state topology may affect the activation entropy $\Delta S^{\ddagger}$. Conformational $\Delta S^{\ddagger}$ is of central interest because its contribution is dominant over that of the energetic contribution in predicting the correct sign of the topology–rate trend in our model (figure 5). Furthermore, a significant degree of correlation between $\Delta S^{\ddagger}$ and the topology is expected because intrachain contacts are constraints on conformational entropy [9, 10]. Here, the transition state topological parameters $CO^{\ddagger}$, $LRO^{\ddagger}$ and $\Lambda_D^{\ddagger}$ are computed by averaging the values taken by the CO, LRO and Λ_D function over the conformations in the interval of Q that defines the transition state ensemble. Several features emerging from these results are noteworthy. First, the correlation is good between the native topological parameters and their transition state counterparts (figures 8(a), 9(a) and 10(a)), indicating that the transition states of the 13 model proteins share similar degrees of similarity with their respective native states. Second, by virtue of this good correlation, the correlation between native topology and folding rate and $\Delta G^{\ddagger}$ (cf figure 4), at least in our model, may be viewed as arising more fundamentally from the correlation between transition state topological parameters and the folding rate. After all, transition state properties and activation quantities are much more directly related to folding rates than native state properties.

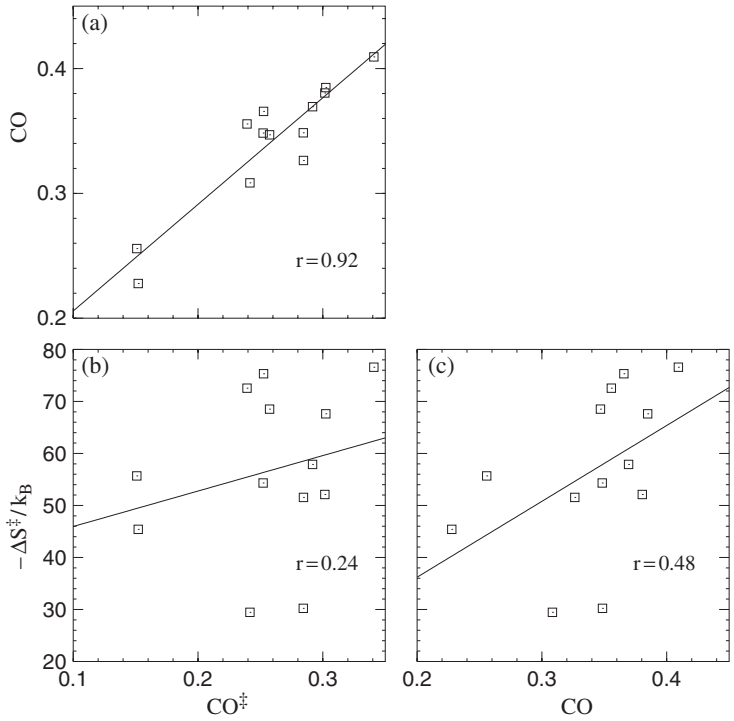

Figure 8. Correlation relationship between the activation entropy $\Delta S^{\ddagger}$ at T_m and the transition state relative contact order $CO^{\ddagger}$ (b), as well as between $\Delta S^{\ddagger}$ and the native relative contact order CO (c) for the set of 13 proteins studied here. (a) shows the corresponding correlation between $CO^{\ddagger}$ and CO.

Third, the correlation between the activation entropy $\Delta S^{\ddagger}$ and the native topological parameters varies, depending on the parameter. The correlation is not so good for CO, reasonable for LRO and quite high for Λ_D (figures 8(c), 9(c) and 10(c)). Because $\Delta S^{\ddagger}$ is a property of the transition state, one intuitively expects a stronger correlation of this quantity with $CO^{\ddagger}$, $LRO^{\ddagger}$ and $\Lambda_D^{\ddagger}$ than with their native counterparts. However, this is not the case for the present model. Figures 8(b), 9(b) and 10(b) show the correlation between $\Delta S^{\ddagger}$ and the transition state topological parameters. The sign of the correlation is consistent with expectation from polymer physics [9, 10], in that higher degrees of topological complexity (higher $CO^{\ddagger}$, $LRO^{\ddagger}$ and $\Lambda_D^{\ddagger}$ values) are associated with lower conformational entropy (higher $-\Delta S^{\ddagger}$ values). But quantitatively these correlations are quite poor except that for $\Lambda_D^{\ddagger}$. They are also significantly weaker than that between $\Delta S^{\ddagger}$ and the corresponding native topological parameters, except again that for $\Lambda_D^{\ddagger}$ ($r = 0.85$) which is comparable to the correlation between $\Delta S^{\ddagger}$ and native Λ_D ($r = 0.88$). These observations imply that the ability to capture conformational entropic effects in folding varies for different topological parameters, and that for the present model Λ_D is enjoying a higher degree of success.

7. Transition state ensembles: explicit-chain model results disagree with topomer search assumptions

The topological parameter Λ_D is a central variable in the topomer search model [28]. Hence, in view of the relative effectiveness of Λ_D in folding rate correlation and in capturing essential conformational entropic properties in the folding of the present set of model proteins, it is

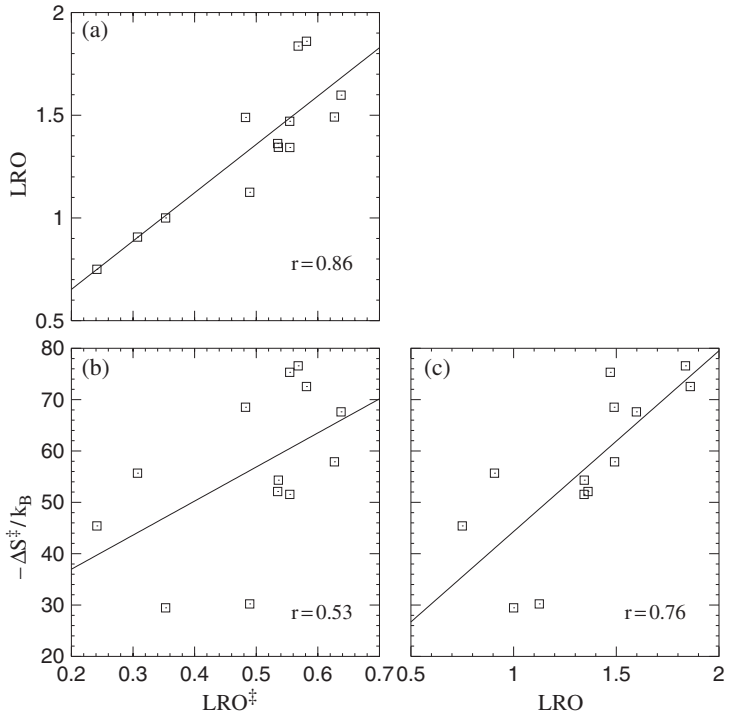

Figure 9. Same as figure 8, but for the transition state LRO‡ and native LRO.

instructive to compare the transition state ensembles postulated by the topomer search model with the ones obtained in the present explicit-chain model. Such an exercise is particularly valuable for addressing questions about conformational entropy, now that a dominant role of conformational entropy is identified in our model. Conformational entropy plays an even more critical role in the topomer search model, which postulates that the rate-limiting step of folding is an essentially unbiased search for the native topomer. This state acts like a transition state of the model, as it is asserted that folding will proceed quickly to the native state once the native topomer has been located [28, 40]. Accordingly, the difference in conformational entropy between the native topomer state and the denatured state is the determining factor for the folding rate in the topomer search model.

However, despite the prominent roles of conformational entropy in both theoretical pictures, the transition states predicted by the topomer search model and the present explicit-chain model are far from similar [41]. Here, this fact is illustrated by the depiction in figure 11 of the transition state ensembles of the explicit-chain model and the native topomers for four proteins, selected for a wide spread in their folding rates. The explicit-chain transition states are much more diffuse than the native topomers, underscoring that the physics of the two theoretical constructs are fundamentally different. Indeed, this trend was already evident from figure 10(a), which shows that although $\Lambda_D^{\ddagger}$ is well correlated with Λ_D, the range of $\Lambda_D^{\ddagger}$ values is only about one third that of Λ_D, implying that the number of nonlocal contacts in the explicit-chain transition state is much smaller than that in the native topomer. Apparently, the entropy–energy compensation in the explicit-chain model is conducive to transition states with significant conformational diversity. The topomer search narrative, on the other hand, was based on a Levinthal-like conformational search process that largely neglected the energetic

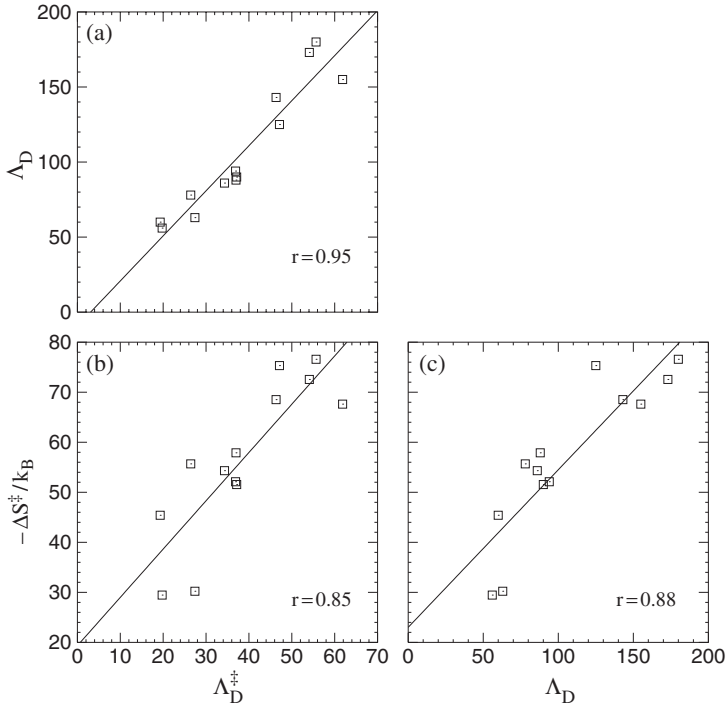

Figure 10. Same as figure 8, but for the transition state $\Lambda_D^{\ddagger}$ and native Λ_D.

contribution. As we have recently discussed, such a folding mechanism is highly unlikely to succeed [41].

8. Experimental folding rates: 'representative' rates from mutational analysis versus actual mid-point rates

Finally, we comment briefly on a procedure recently proposed by Chavez *et al* [48] for extracting an extrapolated zero-stability 'representative' folding rate for a given protein from a large set of folding/unfolding rates of the wild-type protein's single-point mutants. This procedure was motivated by the wish to have a set of experimental folding rates determined at each protein's transition temperature T_m to compare with simulated rates at model T_ms, as discussed in section 4 above, but in many cases the actual folding rate at the heat-denatured transition mid-point in the absence of denaturant has not been measured. The procedure entails extrapolating folding and unfolding rates of a set of single-point mutants measured at a given temperature with zero denaturant, as a function of their native stability $\Delta G_{U-F}^{H_2O}$ (Brønsted plot), to a point at which $\Delta G_{U-F}^{H_2O} = 0$. In this sense, the resulting extrapolated folding rate may be viewed as that of a hypothetical mutant that populates the native and denatured states equally at the given temperature in the absence of denaturant, similar to the behaviour of a protein at its transition mid-point temperature. Chavez *et al* stipulate that the product of this procedure is 'the most representative rate obtainable from experimental data for a given protein structure' (supporting information for [48], page S3). In their analysis, these representative rates are used on the same footing as actual experimentally determined rates at T_ms to assess the agreement between theory and experiment.

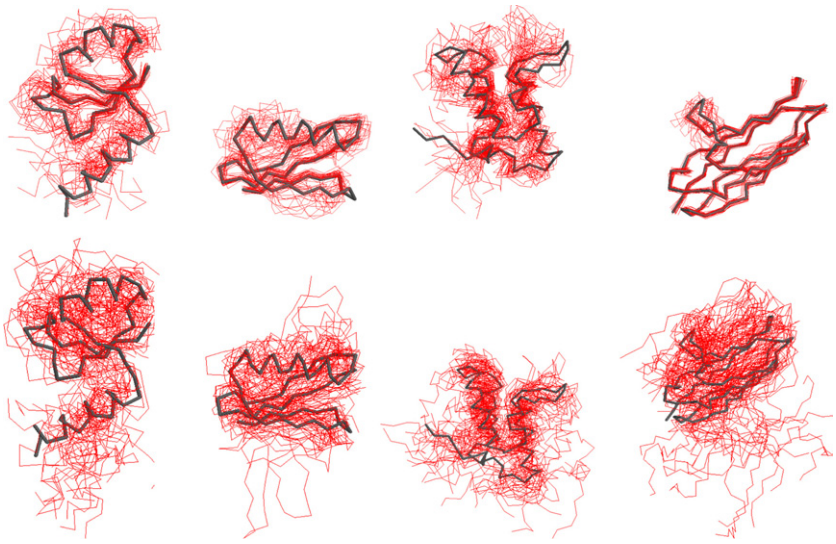

Figure 11. Theoretical transition state ensembles of four of the proteins studied in this work: (*from left to right*) 1div, 1pgb, 1imq and 1wit. Thick black traces depict the backbones of native PDB structures; thin red traces show 25 representative transition state conformations optimally superimposed on each native structure, constructed using the method in [41]. The transition state ensembles postulated by the topomer search model (top row) are compared to that predicted by the present explicit-chain native-centric model (bottom row).

Figure 12 shows the effects of applying this procedure to the data considered in this study. For this purpose, our consideration is restricted to eight proteins that have also been treated by Chavez *et al*. Interestingly, even before using the mid-point and representative rates, merely restricting our data set of thirteen proteins to these eight proteins significantly improves the correlation (from $r = 0.69$ to 0.88) between the simulated folding rates at model T_ms and experimental rates determined near room temperature. This can be attributed in large measure to the fact that the protein 1urn, which is an outlier in figure 2, is not present in the restricted set. When the mid-point and 'representative' rates are used, the correlation of the simulated model T_m rates with the five experimentally determined mid-point rates and three extrapolated representative rates improves slightly to $r = 0.93$, providing some—albeit not conclusive—evidence that simulated T_m rates of the present model match better with experimental mid-point rates than with rates measured near room temperature.

This consideration indicates that the Brønsted procedure of Chavez *et al* is promising. However, several pertinent issues remain to be better elucidated. These include primarily the physical, operational meaning of the representative rate, and its relationship with the actual folding/unfolding rate at the transition mid-point temperature. Protein folding rates are strongly temperature dependent [77–79, 85, 86]. Thus, the relationship between mutant folding rates collected at one temperature and the folding rate of the wild-type protein measured at a higher transition mid-point temperature can be rather complicated. The potential complexity of these issues is highlighted by the comparison in figure 13, which shows a notable difference between the actual folding rate of wild-type 2ci2 at its transition mid-point of 88 °C [78] and the representative rate extrapolated from data collected at 25 °C. This observation means that, in general, the representative rate may not be identified with the wild-type mid-point rate.

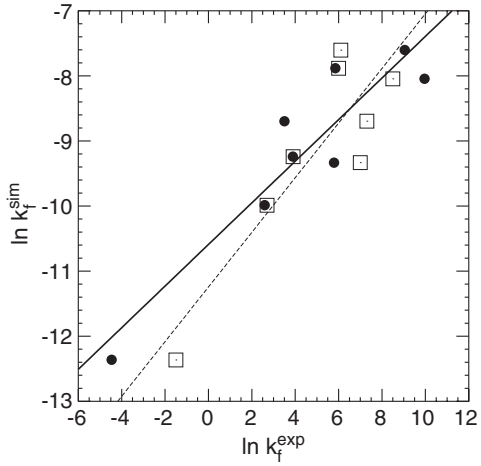

Figure 12. Comparison between the theoretical folding rates of the present native-centric C_α model, $\ln k_f^{sim}$, and two different sets of experimental or experimentally derived folding rates. The set of proteins used in this comparison is the overlap of eight proteins (1aps, 2ci2, 1lmb, 1pgb, 1csp, 1div, 1imq and 1poh) between our data set and the set used by Chavez *et al* [48]. Open squares denote the experimental folding rates near room temperature in table 1 (solid fitted line, $r = 0.88$ for the $\ln k_f^{sim}$–$\ln k_f^{exp}$ correlation). Filled circles denote either the mid-point temperature (2ci2, 1lmb, 1csp, 1div and 1poh) or the 'representative' (1aps, 1pgb and 1imq) folding rates used by Chavez *et al* [48] (dotted fitted line, $r = 0.93$).

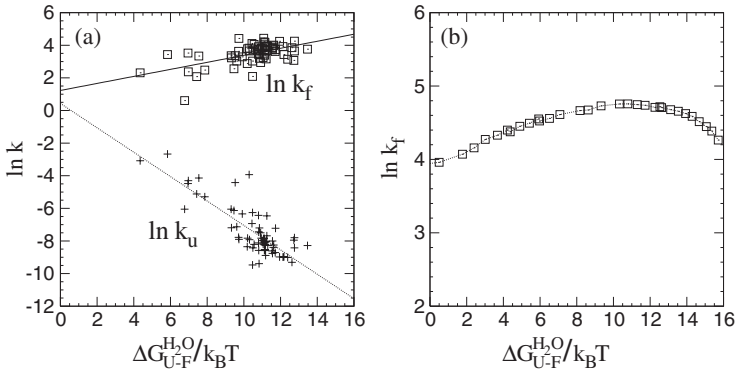

Figure 13. (a) Brønsted plot showing the folding (open squares) and unfolding (crosses) rates for a set of 65 single-point mutants of the protein 2ci2, all measured at $T = 25\,^{\circ}\mathrm{C}$ and zero denaturant concentration, as a function of their native stability $\Delta G_{U-F}^{H_2O}$ (data from [106]). The two linear fits are close to converging at $\Delta G_{U-F}^{H_2O} = 0$, at an extrapolated rate $\ln k_0^{exp} \approx 1$. (b) Actual experimental temperature dependence of the folding rate $\ln k_f^{exp}$ of wild-type 2ci2 as a function of native stability (rate stability plot adapted from [85], original experimental data from [78, 88]). The mid-point temperature folding rate is found at $\Delta G_{U-F}^{H_2O} = 0$; this rate differs significantly from $\ln k_0^{exp}$.

9. Concluding remarks and outlook

In summary, the main findings of our modelling exercise are as follows. The dominant role of conformational entropic barrier in begetting the topology–rate relationship in the present class of native-centric explicit-chain protein models suggests that similar conformational entropic

effects may be at play in the topology-dependent folding of real proteins. For our models, we find that the topomer search model parameter Λ_D and its transition state counterpart $\Lambda_D^{\ddagger}$ are particularly effective in capturing the pertinent conformational activation entropies. In this light, it is likely that much remains to be learnt from the predictive power of these and the related long range order parameters, even though the original physical interpretation of the Λ_D parameter [28] is problematic [41].

It has been pointed out that the success of a native topological parameter in folding rate correlation, such as that of the original CO, does not by itself prove that the parameter is directly related to the underlying folding mechanism, because the given parameter can be 'a proxy for some other, physically more reasonable parameter' [28, 107]. In the context of the discussion in which this possibility was raised, the 'physically more reasonable parameter' was envisioned to be a differently defined topological variable for the native state [28, 107]. The present investigation goes one step further. Our analysis makes it clear that, more fundamentally, all native topological parameters may well be 'proxies' for certain rate-determining properties of the transition state. For instance, Λ_D may well be a proxy for $\Lambda_D^{\ddagger}$, which in turn may be viewed as a proxy for $\Delta S^{\ddagger}$, etc.

As demonstrated by the results of our study, this generalization of the 'proxy' idea in topology–rate relationship serves to open up new avenues for establishing physical connections from the empirical native topological parameters to the actual folding process. However, more in-depth analyses will be necessary to pin down the energetics that underlies the new statistical correlations that we have observed. A focus on activation barriers to protein folding also necessitates a better delineation of the role of the denatured states. In this regard, it would be useful to compare the present approach with other explicit-chain methods, such as a recently proposed 'route measure' [48], that also aim to address the kinetic effects of conformational entropy. Although key advances have been made using the present class of coarse-grained native-centric models, these constructs are limited in several important respects, leaving room for improvement, for example, in their treatment of desolvation effects [63, 84, 108] and enthalpic protein folding barriers [84–86]. Thus, inferences from these models for real proteins should be considered tentative. It would be instructive to ascertain whether the application of improved self-contained explicit-chain models that incorporate these features would lead to a better mimicry of the experimental topology–rate trend.

Acknowledgments

This work was partly supported by the Swedish Research Council (SW) and a Canadian Institutes of Health Research (CIHR) grant to HSC, who holds a Canada Research Chair in Proteomics, Bioinformatics and Functional Genomics.

References

[1] Plaxco K W, Simons K T and Baker D 1998 *J. Mol. Biol.* **277** 985
[2] Jackson S E 1998 *Fold. Des.* **3** R81
[3] Chan H S 1998 *Nature* **392** 761
[4] Bryngelson J D, Onuchic J N, Socci N D and Wolynes P G 1995 *Proteins Struct. Funct. Genet.* **21** 167
[5] Thirumalai D and Woodson S A 1996 *Acc. Chem. Res.* **29** 433
[6] Dill K A and Chan H S 1997 *Nat. Struct. Biol.* **4** 10
[7] Mirny L and Shakhnovich E 2001 *Annu. Rev. Biophys. Biomol. Struct.* **30** 361
[8] Chan H S, Shimizu S and Kaya H 2004 *Methods Enzymol.* **380** 350
[9] Chan H S and Dill K A 1989 *J. Chem. Phys.* **90** 492
Chan H S and Dill K A 1992 *J. Chem. Phys.* **96** 3361 (erratum)
Chan H S and Dill K A 1997 *J. Chem. Phys.* **107** 10353 (erratum)

[10] Chan H S and Dill K A 1990 *J. Chem. Phys.* **92** 3118
Chan H S and Dill K A 1997 *J. Chem. Phys.* **107** 10353 (erratum)
[11] Chan H S and Dill K A 1990 *Proc. Natl Acad. Sci. USA* **87** 6388
[12] Tadokoro H 1979 *Structure of Crystalline Polymers* (New York: Wiley)
[13] Maritan A, Micheletti C, Trovato A and Banavar J R 2000 *Nature* **406** 287
[14] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2004 *Proc. Natl Acad. Sci. USA* **101** 7960
[15] Yee D P, Chan H S, Havel T F and Dill K A 1994 *J. Mol. Biol.* **241** 557
[16] Hubner I A and Shakhnovich E I 2005 *Phys. Rev.* E **72** 022901
[17] Kim N, Shiffeldrim N, Gan H H and Schlick T 2004 *J. Mol. Biol.* **341** 1129
[18] Gonzalez O and Maddocks J H 1999 *Proc. Natl Acad. Sci. USA* **96** 4769
[19] Buck G R and Zechiedrich E L 2004 *J. Mol. Biol.* **340** 933
[20] Flammini A, Maritan A and Stasiak A 2004 *Biophys. J.* **87** 2968
[21] Taylor W R, Xiao B, Gamblin S J and Lin K 2003 *Comput. Biol. Chem.* **27** 11
[22] Zhou H-X 2004 *Acc. Chem. Res.* **37** 123
[23] Mallam A L and Jackson S E 2005 *J. Mol. Biol.* **346** 1409
[24] Selvaraj S and Gromiha M M 2001 *J. Mol. Biol.* **310** 27
[25] Zhou H and Zhou Y 2002 *Biophys. J.* **82** 458
[26] Micheletti C 2003 *Proteins Struct. Funct. Genet.* **51** 74
[27] Gong H, Isom D G, Srinivasan R and Rose G D 2003 *J. Mol. Biol.* **327** 1149
[28] Makarov D E and Plaxco K W 2003 *Protein Sci.* **12** 17
[29] Ivankov D N and Finkelstein A V 2004 *Proc. Natl Acad. Sci. USA* **101** 8942
[30] Punta M and Rost B 2005 *J. Mol. Biol.* **348** 507
[31] Baker D 2000 *Nature* **405** 39
[32] Alm E and Baker D 1999 *Proc. Natl Acad. Sci. USA* **96** 11305
[33] Muñoz V and Eaton W A 1999 *Proc. Natl Acad. Sci. USA* **96** 11311
[34] Galzitskaya O V and Finkelstein A V 1999 *Proc. Natl Acad. Sci. USA* **96** 11299
[35] Weikl T R and Dill K A 2003 *J. Mol. Biol.* **329** 585
[36] Kaya H and Chan H S 2000 *Proteins: Struct. Funct. Genet.* **40** 637
[37] Karanicolas J and Brooks C L III 2003 *Proteins: Struct. Funct. Genet.* **53** 740
[38] Debe D A, Carlson M J and Goddard W A III 1999 *Proc. Natl Acad. Sci. USA* **96** 2596
[39] Debe D A and Goddard W A III 1999 *J. Mol. Biol.* **294** 619
[40] Makarov D E, Keller C A, Plaxco K W and Metiu H 2002 *Proc. Natl Acad. Sci. USA* **99** 3535
[41] Wallin S and Chan H S 2005 *Protein Sci.* **14** 1643
[42] Zhou H-X 2005 *Phys. Biol.* **2** R1
[43] Koga N and Takada S 2001 *J. Mol. Biol.* **313** 171
[44] Faisca P F N and Ball R C 2002 *J. Chem. Phys.* **117** 8587
[45] Cieplak M and Hoang T X 2003 *Biophys. J.* **84** 475
[46] Jewett A I, Pande V S and Plaxco K W 2003 *J. Mol. Biol.* **326** 247
[47] Kaya H and Chan H S 2003 *Proteins Struct. Funct. Genet.* **52** 524
[48] Chavez L L, Onuchic J N and Clementi C 2004 *J. Am. Chem. Soc.* **126** 8426
[49] Ejtehadi M R, Avall S P and Plotkin S S 2004 *Proc. Natl Acad. Sci. USA* **101** 15088
[50] Faisca P F N, da Gama M M T and Nunes A 2005 *Proteins Struct. Funct. Bioinform.* **60** 712
[51] Fersht A R 1995 *Proc. Natl Acad. Sci. USA* **92** 10869
[52] Du R, Pande V S, Grosberg A Y, Tanaka T and Shakhnovich E I 1998 *J. Chem. Phys.* **108** 334
[53] Bilsel O and Matthews C R 2000 *Adv. Protein Chem.* **53** 153
[54] Nymeyer H, Socci N D and Onuchic J N 2000 *Proc. Natl Acad. Sci. USA* **97** 634
[55] Kaya H and Chan H S 2002 *J. Mol. Biol.* **315** 899
[56] Hubner I A, Shimada J and Shakhnovich E I 2004 *J. Mol. Biol.* **336** 745
[57] Bai Y, Zhou H and Zhou Y 2004 *Protein Sci.* **13** 1173
[58] Day R and Daggett V 2003 *Adv. Protein Chem.* **66** 373
[59] Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu K Y and García A E 2003 *Curr. Opin. Struct. Biol.* **13** 168
[60] Snow C D, Sorin E J, Rhee Y M and Pande V S 2005 *Annu. Rev. Biophys. Biomol. Struct.* **34** 43
[61] Das P, Matysiak S and Clementi C 2005 *Proc. Natl Acad. Sci. USA* **102** 10141
[62] Clementi C, Nymeyer H and Onuchic J N 2000 *J. Mol. Biol.* **298** 937
[63] Kaya H and Chan H S 2003 *J. Mol. Biol.* **326** 911
Kaya H and Chan H S 2004 *J. Mol. Biol.* **337** 1069 (corrigendum)
[64] Taketomi H, Ueda Y and Gō N 1975 *Int. J. Pept. Protein Res.* **7** 445
[65] Micheletti C, Banavar J R, Maritan A and Seno F 1999 *Phys. Rev. Lett.* **82** 3372

[66] Shea J-E, Onuchic J N and Brooks C L III 1999 *Proc. Natl Acad. Sci. USA* **96** 12512
[67] Kaya H and Chan H S 2000 *Phys. Rev. Lett.* **85** 4823
[68] Knott M, Kaya H and Chan H S 2004 *Polymer* **45** 623
[69] Das P, Wilson C J, Fossati G, Wittung-Stafshede P, Matthews K S and Clementi C 2005 *Proc. Natl Acad. Sci. USA* **102** 14569
[70] Zuo G H, Zhang J, Wang J and Wang W 2005 *Chin. Phys. Lett.* **22** 1809
[71] Kamagata K, Arai M and Kuwajima K 2004 *J. Mol. Biol.* **339** 951
[72] Berendsen H J, Postma J P M, van Gunsteren W F, DiNola A and Haak J R 1984 *J. Chem. Phys.* **81** 3684
[73] Veitshans T, Klimov D and Thirumalai D 1997 *Fold. Des.* **2** 1
[74] Guo Z and Thirumalai D 1995 *Biopolymers* **36** 83
[75] Allen M P and Tildesley D J 1987 *Computer Simulations of Liquids* (Oxford: Oxford University Press)
[76] Zuckerman D M and Woolf T B 2002 *J. Chem. Phys.* **116** 2586
[77] Segawa S-I and Sugihara M 1984 *Biopolymers* **23** 2473
[78] Oliveberg M, Tan Y-J and Fersht A R 1995 *Proc. Natl Acad. Sci. USA* **92** 8926
[79] Scalley M L and Baker D 1997 *Proc. Natl Acad. Sci. USA* **94** 10636
[80] Chan H S 1998 *Monte Carlo Approach to Biopolymers and Protein Folding* ed P Grassberger, G T Barkema and W Nadler (Singapore: World Scientific) p 29
[81] Chan H S and Dill K A 1998 *Proteins: Struct. Funct. Genet.* **30** 2
[82] Salvi G and De Los Rios P 2003 *Phys. Rev. Lett.* **91** 258102
[83] Kaya H and Chan H S 2003 *Proteins: Struct. Funct. Genet.* **52** 510
[84] Kaya H, Liu Z and Chan H S 2005 *Biophys. J.* **89** 520
[85] Liu Z and Chan H S 2005 *J. Mol. Biol.* **349** 872
[86] Liu Z and Chan H S 2005 *Phys. Biol.* **2** S75
[87] Ivankov D N, Garbuzynskiy S O, Alm E, Plaxco K W, Baker D and Finkelstein A V 2003 *Protein Sci.* **12** 2057
[88] Jackson S E and Fersht A R 1991 *Biochemistry* **30** 10428
[89] Van Nuland N A, Chiti F, Taddei N, Rauei G, Ramponi G and Dobson C M 1998 *J. Mol. Biol.* **283** 883
[90] Silow M and Oliveberg M 1997 *Biochemistry* **36** 7633
[91] Burton R E, Huang G S, Daugherty M A, Fullbright P W and Oas T G 1996 *J. Mol. Biol.* **263** 311
[92] Plaxco K W, Guijarro J I, Morton C J, Pitkeathly M, Campbell I D and Dobson C M 1998 *Biochemistry* **37** 2529
[93] McCallister E L, Alm E and Baker D 2000 *Nat. Struct. Biol.* **7** 669
[94] Clarke J, Cota E, Fowler S B and Hamill S J 1999 *Struct. Fold. Des.* **7** 1145
[95] Schindler T, Herrler M, Maraheil M A and Schmid F X 1995 *Nat. Struct. Biol.* **2** 663
[96] Perl D, Welker C, Schindler T, Schroder K, Marahiel M A, Jaenicke R and Schmid F X 1998 *Nat. Struct. Biol.* **5** 229
[97] Otzen D E and Oliveberg M 1999 *Proc. Natl Acad. Sci. USA* **96** 11746
[98] Kuhlman B, Luisi D L, Evans P A and Raleigh D P 1998 *J. Mol. Biol.* **284** 1661
[99] Ferguson N, Capaldi A P, James R, Kleanthous C and Radford S E 1999 *J. Mol. Biol.* **286** 1597
[100] Van Nuland N A, Meijberg W, Warner J, Forge V, Scheek R M, Robillard G T and Dobson C M 1998 *Biochemistry* **37** 622
[101] Li M S, Klimov D K and Thirumalai D 2004 *Polymer* **45** 573
[102] Kaya H and Chan H S 2003 *Phys. Rev. Lett.* **90** 258104
[103] Fersht A R, Matouschek A and Serrano L 1992 *J. Mol. Biol.* **224** 771
[104] Hubner I A, Oliveberg M and Shakhnovich E I 2004 *Proc. Natl Acad. Sci. USA* **101** 8354
[105] Kramer H A 1940 *Physica* **7** 284
[106] Itzhaki L S, Otzen D E and Fersht A R 1995 *J. Mol. Biol.* **254** 260
[107] Gillespie B and Plaxco K W 2004 *Annu. Rev. Biochem.* **73** 837
[108] Cheung M S, García A E and Onuchic J N 2002 *Proc. Natl Acad. Sci. USA* **99** 685
[109] Micheletti C 2002 personal communication