doi:10.1006/jmbi.2001.5037 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **313**, 171–180

JMB

Roles of Native Topology and Chain-length Scaling in Protein Folding: A Simulation Study with a Gō-like Model

Nobuyasu Koga[1] and Shoji Takada[1,2]*

[1] *Graduate School of Science and Technology, Kobe University and*

[2] *PRESTO, Japan Science and Technology, Rokkodai, Nada Kobe 657-8501, Japan*

**Corresponding author*

We perform folding simulations on 18 small proteins with using a simple Gō-like protein model and analyze the folding rate constants, characteristics of the transition state ensemble, and those of the denatured states in terms of native topology and chain length. Near the folding transition temperature, the folding rate k_F scales as $k_F \sim \exp(-c\ RCO\ N^{0.6})$ where *RCO* and *N* are the relative contact order and number of residues, respectively. Here the topology *RCO* dependence of the rates is close to that found experimentally ($k_F \sim \exp(-c\ RCO)$), while the chain length *N* dependence is in harmony with the predicted scaling property ($k_F \sim \exp(-c\ N^{2/3})$). Thus, this may provides a unified scaling law in folding rates at the transition temperature, $k_F \sim \exp(-c\ RCO\ N^{2/3})$. The degree of residual structure in the denatured state is highly correlated with *RCO*, namely, proteins with smaller *RCO* tend to have more ordered structure in the denatured state. This is consistent with the observation that many helical proteins such as myoglobin and protein A, have partial helices, in the denatured states. The characteristics of the transition state ensemble calculated by the current model, which uses native topology but not sequence specific information, are consistent with experimental ϕ-value data for about half of proteins.

Keywords: contact order; folding rate; transition state; ϕ-value; denatured state

Introduction

How proteins fold to their native structures has long been studied and still remains a fundamental question in biophysics.[1–3] Recent studies of small, single-domain proteins have leaded to a new paradigm that the folding rates and pathways are largely determined by their native topology. First, Baker and co-workers found that the folding rates of two-state folding proteins are strongly correlated with a measure of the topological complexity of native structure,[4,5] that is, folding is slower for more complex topologies. The measure of topological complexity, which they called the relative contact order, is a key quantity currently. It is defined as the average sequence separation of native contacts, which are the residue pairs close in contact at the native structure, divided by the total number of residues. Second, the folding transition state ensembles (TSEs) are characterized by the ϕ-value analysis developed by Fersht *et al.*[1,6] for many proteins and it was found that proteins that have similar structures with low sequence homology have similar TSE for the SH3 domain family[7,8] and αβ plaits[9,10] which implies that the native topology tends to determine folding pathways of the protein. It should be noted, however, those recent works report some exceptions to this scenario:[11–13] For example, protein G and protein L have the same topology, but exhibit different transition state ensembles.

Why native topology plays an important role in determining folding mechanisms is qualitatively understood from the energy landscape perspective. It suggests that fast folding proteins have a funnel-like energy landscape with some ruggedness on the surface.[2] Evolutionary pressure has made the ruggedness on the slope sufficiently small so that the current natural proteins may have minimal

Abbreviations used: TSE, transition state ensembles; *RCO*, relative contact order; *ACO*, absolute contact order; D, denatured; N, native; AcP, acylphosphatase; ADA2h, procarboxypeptidase A2.

E-mail address of the corresponding author: stakada@kobe-u.ac.jp

0022-2836/01/010171–10 $35.00/0

frustration. The effective energy, which is defined as protein internal interaction energy plus solvation free energy, decreases smoothly with folding reaction to proceed, whereas the acquisition of structured segments upon folding reduces the chain entropy, too. Since the energy stabilization accelerates and the entropy loss decelerates folding reaction, their subtle balance determines the folding pathways and rates. Upon formation of a residue-residue contact, the entropy loss is larger for non-local contact than for local one. Thus, a protein whose native topology has more non-local contacts folds more slowly. Also, qualitatively, contacts formed at the TSE are chosen so that the entropy loss may be minimized with given amount of structured segments. Here the entropy loss is strongly dependent on the sequence separation of the formed contact and so the native topology controls characteristics of the folding TSEs.

If we move from the qualitative to the semi-quantitative level, however, many issues need to be clarified. First, it is somewhat surprising that the folding rate constant correlates with relative contact order instead of absolute contact order[†] because the latter is more directly linked to the entropy loss upon contact formation.[3] Related to this is chain-length scaling of the folding rate, where theoretical work predicts some different types of scaling such as $k_F \sim N^{-\nu}$,[14] $k_F \sim \exp(-cN^{2/3})$,[15] $k_F \sim \exp(-cN^{1/2})$,[16] that were unified by Wolynes[17] while experimental data exhibit little correlation between chain lengths and folding rates.[5] Here, experimental data are collected only for small two-state folding proteins, while the theory dealt with relatively large proteins and thus this may not be inconsistency, but a gap between small and large proteins. Here, we show how the contact order and chain length scaling can be coupled to fill the gap giving a unified perspective of folding rates.

Another challenge at the semi-quantitative level is the prediction of folding pathways and characteristics of the TSE from a simple funnel theory. In particular, comparison of the calculated ϕ values with experimental ones is an excellent way of testing the funnel theory. During the last few years, several groups reported that simple free energy functionals can describe the folding free energy surface and quantitatively predict folding pathways and rate constants with reasonable accuracy.[18–23] A basic assumption common in all of these works is that the native contacts (the contacts that exist in the native structure) alone can describe the overall shape of a funnel energy landscape and interaction energies of any non-native contacts are neglected. These models are often called perfect-funnel models, or Gō-like models.[24,25] The functional model essentially has two competing contributions, one from energetic stability due to contacts, and the other from entropy change due to loss of chain flexibility.

Advantages of these free-energy functional approaches are their simplicity and reasonable agreement between theory and experiments. On the other hand, a possible shortcoming may be the difficulty to single out reasons for any discrepancy between theory and experiments because these theories often are based on many drastic approximations, such as neglect of non-native interactions, crude approximations in the expression of entropy loss and sequence-specific interactions included in the model.

An alternative and complementary approach to these studies was recently taken by Onuchic and his colleagues.[26,27] They proposed a simple off-lattice simulation model that implements Gō-like interactions and performed molecular dynamics simulation for a few small-to-medium size proteins. They reported probabilities of forming native contacts in the TSEs and intermediate states that are qualitatively consistent with experiments. This simulation study seems more straightforward than the free-energy functional approaches. Especially, no additional approximation is needed besides the interaction energy function. It is also based on the perfect-funnel model and is constructed mostly from the native topology without taking into account details of sequence-specific information. Thus, it is less ambiguous to address the interplay between the native topology and folding mechanisms, making it easier to single out reasons for a discrepancy, if any, with experiments.

Here, we attempt to extend Clementi, Nymeyer, and Onuchic's work[26] to a more comprehensive level. Performing molecular dynamics simulation on 18 small proteins, most of which are experimentally known to exhibit apparent two-state transition, we address in greater detail the relation between folding mechanisms and the native topology. In particular, the folding rates are computed and their correlation with the contact order and its variants is investigated; the degree of native-like order in the denatured states is analyzed in terms of the contact order; and the folding pathways are analyzed and computed TSEs, i.e. computed ϕ-values, are compared with experimental ϕ-value results.

Results

We use a minimal protein model proposed by Clementi *et al*.[26,27] that tries to mimic the perfect funnel aspect of folding energy landscape. Briefly, the protein chain is represented only with C^{α} atoms of every amino acid residues that are connected *via* virtual bonds. Both local and non-local interactions are set up so that they have the lowest energy at the native structure. In particular, attractive contact interactions are introduced between native contacts, which are the residue pairs in close contact in the native structure, while the rest of

[†] The absolute contact order is the average sequence separation of native contacts not divided by the chain length.

pairs have solely repulsive interactions avoiding chain crossing.

During folding simulation at a temperature T, we monitor a measure of nativeness, Q-score. The value of Q close to unity means the conformation is similar to the native structure whilst a Q-value of zero means the conformation is dissimilar to the native structure. At a sufficiently high (low) temperature, Q is mostly smaller (larger) than 0.5. In between them, we can find a temperature where $Q < 0.5$ for half of time and $Q > 0.5$ for the rest of time (see Figure 1(a) for the case of src SH3 domain), which should be close to the transition temperature. Folding and unfolding rate constants are strongly temperature dependent and thus it is crucial to determine the transition temperature accurately. To this end, with use of the above mentioned simulation, we employ the multiple histogram method,[28] computing the heat capacity ($C_v = (\langle E^2\rangle - \langle E\rangle^2)/kT^2$) and change in the Q score ($-\,d\langle Q\rangle/dT = (\langle EQ\rangle - \langle E\rangle\langle Q\rangle)/kT^2$) as a function of temperature, T. (see Figure 1(b) for C_v of the src SH3 domain) Both curves have peaks at the same temperature which we define as the folding transition temperature T_F. We calculated the free energy profile as a function of Q by the histogram method, confirming the two-state transition (see Figure 1(c) for the case of src SH3 domain).

In the following sub-sections, we analyze folding mechanisms for 18 single-domain proteins, most of which are experimentally considered to exhibit the two-state folding transition. The proteins studied include mainly α, mainly β, and α/β mixed proteins ranging from 53 to 153 residues long.

Folding rate constants

We have computed the folding rate constants at $T = 0.97T_F$ for 18 small proteins. Figure 2(a) depicts the folding rate constants (in a logarithmic scale) as a function of the chain length, N, where we see little correlation. This is consistent with what Baker *et al*. found in experimental data.[5] On the other hand, Figure 2(b) and (c) show the folding rates

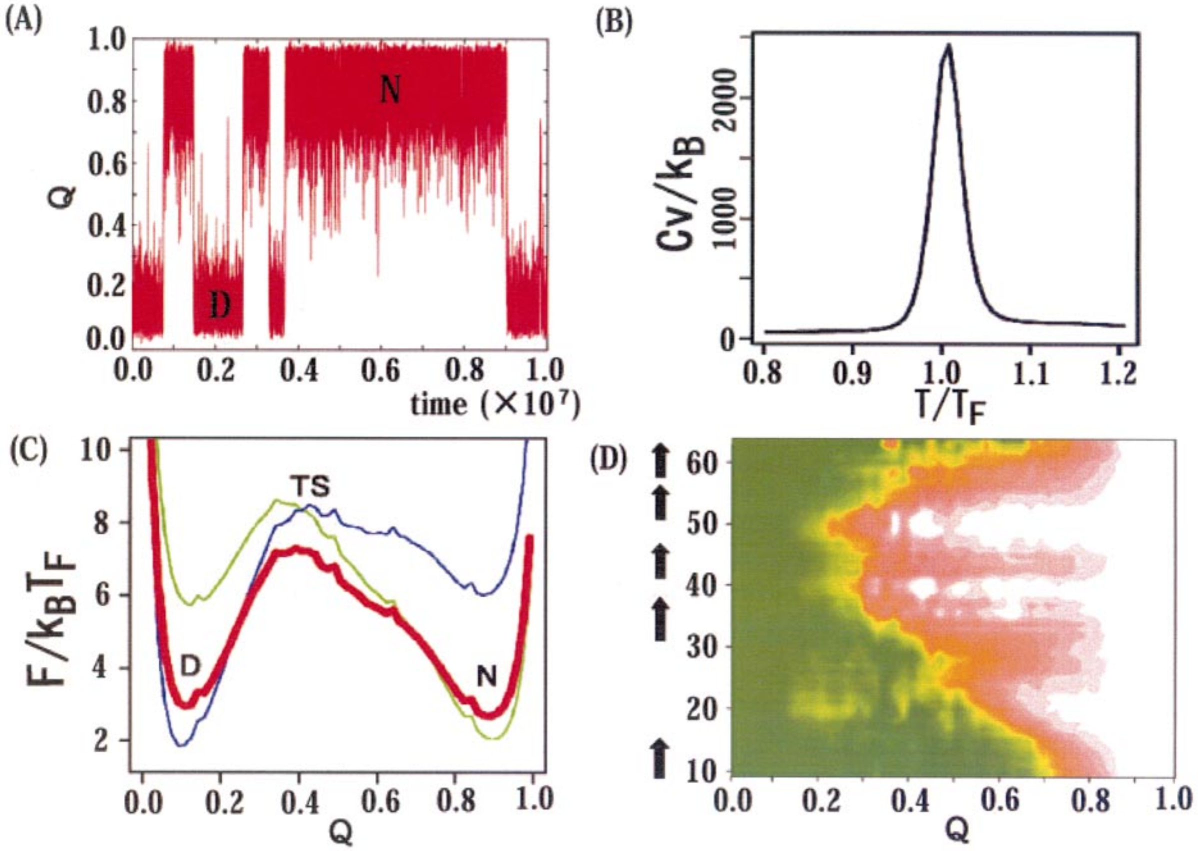

Figure 1. Example of folding time course and statistical mechanical analysis for the case of src SH3 domain. (a) The native-ness Q-value as a function of time t near the folding transition temperature T_F. This shows apparent two-state transition between native ($Q \sim 0.9$) and denatured ($Q \sim 0.15$) states. (b) C_v as a function of temperature T, which exhibits a clear peak at the folding transition temperature T_F. (c) The free energy profiles $F(Q)$ near the folding transition temperature T_F, $T = 0.96T_F$ (green), $T = T_F$ (the thick red curve), and $T = 1.04T_F$ (blue). The denatured and native states are separated by a free energy barrier around $Q \sim 0.4$. (d) The site-resolved folding pathways q_i (Q) (see the text for the explicit definition) are plotted along the reaction coordinate Q. $q_i(Q)$ is unity (zero) when the local environment of site i is native-like (denatured-like) and the value of $q_i(Q)$ is represented by the color; green, yellow, and white correspond to 0, 0.5, and 1, respectively. Positions of β-strands are illustrated as arrows in the left side.

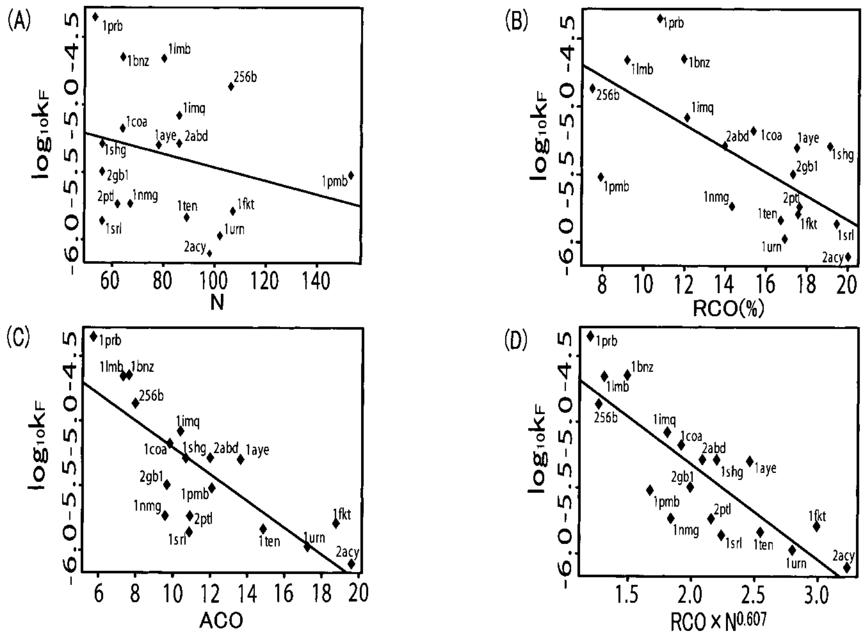

Figure 2. Folding rate constants of 18 small proteins simulated at $T = 0.97T_F$. (a) The rates are plotted against the chain length N where little correlation is seen. (b) The rates are plotted with respect to the relative contact order (RCO). The correlation is found with $r = 0.69$. (c) The rates are plotted against the absolute contact order (ACO) where a stronger correlation ($r = 0.80$) is found. (d) The rates are plotted with respect to RCO x $N^{0.607}$ at which the strongest correlation is obtained ($r = 0.84$).

with respect to the relative contact order (RCO, $RCO = \Sigma S_{ij}/MN$ where S_{ij} is the sequence separation between i and j, the summation is over all native contacts, and M is the number of native contacts) and the absolute contact order (ACO, $ACO = \Sigma S_{ij}/M$), respectively. We find significant correlation in both cases as in experimental data, but the correlation here is somewhat stronger with ACO (the correlation coefficient $r = 0.80$) than with RCO ($r = 0.69$), which is opposite to what was found experimentally.[3] We note that the folding rates are computed near T_F while the experimental data are mostly collected around 20-25 °C, which are lower than T_F. Thus the difference is not necessarily a contradiction, but rather reflects the more controlled conditions made possible in the simulation.

If we hypothetically change the chain length with the topology fixed, we see that the RCO is a scale-invariant quantity and $ACO = RCO \times N$ is proportional to the chain length. To further investigate the chain length scaling, we perform a nonlinear least square fitting of the folding rates ($\ln k_F = A - B \cdot RCO \cdot N^{\nu}$). This yields $\nu = 0.607 \pm 0.179$ with $r = 0.84$. The folding rates are plotted against $RCO \cdot N^{0.607}$ in Figure 2(d). Interestingly, this is in harmony with the scaling property $k_F \sim \exp(-c\ N^{2/3})$ predicted by Finkelstein & Badredtinov[15] and Wolynes[17] at the folding transition temperature.

We also remark that the slope in the plot of ln k_F *versus* RCO (Figure 2(b)) is much less than experimentally observed.[5] The measured folding time varies over six orders of magnitude depending on topology, while simulated ones are within two orders of magnitude. In simulations, the slope depends on energy balance between the local and non-local interactions. By increasing non-local interactions relative to local ones, we get much slower rates for complex proteins with large RCO, while the rate constants of helical proteins are not affected as much. Simply due to limit of computer time, we use non-local interactions somewhat weaker than that should be for the same slope in ln k_F plot. We anticipate that real proteins have stronger non-local interactions than those in this simulation.

Residual structure in the denatured state

Next, we investigate the degree of native-like order in the denatured (D) state, in the transition state ensemble (TSE) as well as in the native (N) state. The native and denatured states are defined as the local minima of the free energy profile (see Figure 1(c) for src SH3 domain), while the TSE corresponds to the barrier top. The Q-value for each of the three states is plotted in Figure 3 against the RCO, where the red, green and blue dots correspond to the D state, the TSE and the N state, respectively.

First, we find that the Q-value of the D state, Q(D), is highly correlated with RCO with a negative correlation coefficient ($r = -0.87$); namely proteins with small RCO, which are mainly composed

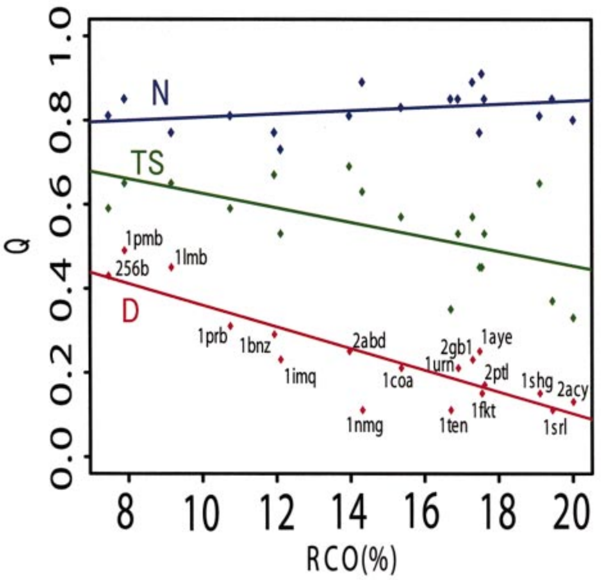

Figure 3. The average native-ness Q-values in the denatured states, the native states, and the TSE are plotted with respect to the relative contact order (RCO) at the folding transition temperature. We see that the degree of residual structures in the denatured states are larger and more significant for α-helical proteins than for β-proteins.

of α-helices, tend to have higher degree of native-like order in the D state. This is consistent with experimental observations. Although it is not easy to quantify the residual order in the D state of real proteins, there are many papers that report residual helical structure in helical proteins. For example, A and H helices of myoglobin[29] and the third helix in the B domain of protein A[30] are considered to be partially formed in the D state. Some helices are inherently stable without tertiary contacts and so some secondary structures may remain in the D state. On the other hand, any β-strand can not be stable of its own, and thus it is relatively rare for partial β-structure survives in the D state. The Q-value of the TSE seems to have the same tendency as that of the D state, although the correlation is weaker.

The Q-values of the N state, Q(N), are around 0.8 for all proteins, almost independent of the RCO. A weak positive correlation is seen in Figure 3, which probably reflects that proteins with higher RCO have a more global network of contacts that reduces the fluctuation in the N state. Fluctuation amplitude in the N state of real proteins, however, should depend on the detailed atomic packing as well as coarse-grained topological network and so the weak dependence found here might be an artifact of the current simple model.

Folding pathway ensemble and topology

To quantify folding pathways, we introduce a local measure of nativeness $q_i(X)$ for the ith amino acid residue at a given ensemble of structures X:

$$q_i(X) = \frac{\langle e_i \rangle_X - \langle e_i \rangle_D}{\langle e_i \rangle_N - \langle e_i \rangle_D}$$

where e_i is the site-energy defined as a sum of interaction energies of ith amino acid residue with any other residues. The bracket $\langle\ \rangle$ means an average of the quantity over an ensemble that is specified as subscript; N and D are the native and denatured states. In particular, $\phi_i = q_i(TS)$ is an analog of experimental ϕ-value of site i, under the assumptions of simple transition state theory, and is used for comparison with protein engineering experiments.

SH3 domain and Sso7d

The SH3 domain is an interesting example to investigate the role of topology versus sequence-specificity for determining folding pathways because comprehensive ϕ-value analysis was performed for two different proteins in the family, src and α-spectrin SH3 domains,[7,8] as well as its structural analog Sso7d.[23] Although the src and α-spectrin SH3 domains have only a weak sequence homology (36 %), their folding transition state ensembles turned out to be largely similar. Namely, the distal β-hairpin and nearby contiguous region are highly structured, while the rest of the protein is less ordered at the transition state ensemble. This similarity to the transition state lead to an idea that the protein topology can determine the folding transition state. Since the current simulation model mostly utilizes topological information of the protein, neglecting sequence specificity, it is interesting to see if the characteristics of the TSE are well reproduced by the current model. Onuchic *et al*. already reported simulation results for src SH3 finding qualitative agreement with experiments,[26] which we now investigate more comprehensively and quantitatively.

Figure 1(d) shows computed folding pathway ensembles of src SH3 domain. In Figure 1(d), local measure of native-ness, $q_i(Q)$ is represented along the global reaction coordinate Q. We see that the distal β-hairpin, which corresponds to the third and fourth β-strands, indeed gets native-like order at an earlier stage than the rest of the protein. This is consistent with the experimental results. More specifically, the TSE corresponds to $Q = 0.4$ where ϕ_i, a computational analog of the experimental ϕ-values, are computed. This is plotted onto the native structure in Figure 4(a, left), together with experimentalf values in Figure 4(a, right). Here, the blue and red stand for unstructured ($\phi = 0$) and structured ($\phi = 1$), at the TSE, respectively. Clearly, the major features of theory and experiment resembles each other albeit minor discrepancy. The

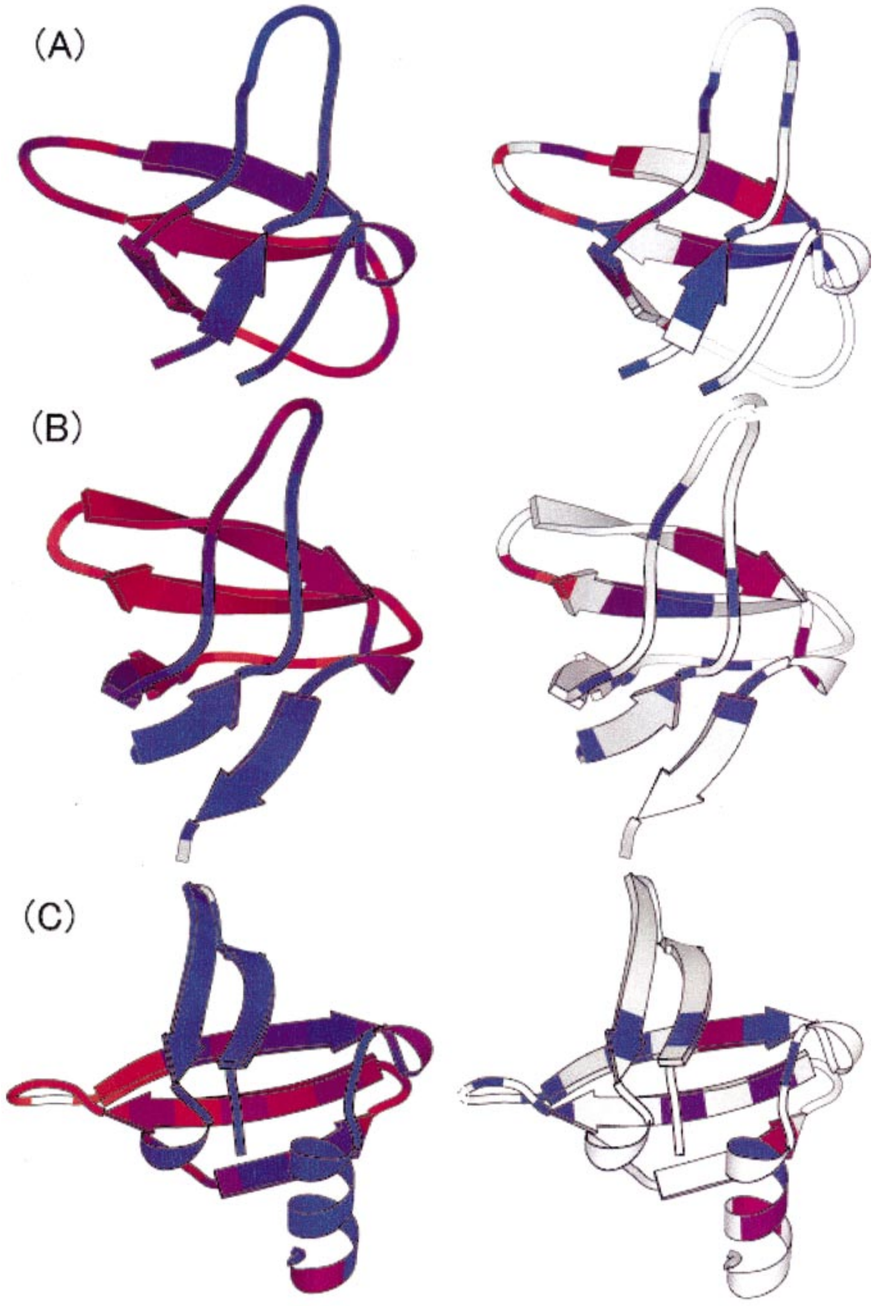

Figure 4. Both experimental and computed ϕ-values of SH3 domains and their structural analog. The blue and red mean $\phi = 0$ (unstructured in the TSE), and 1 (native-like in the TSE), respectively, whereas white segments mean no available experimental data or ϕ-values larger than unity or smaller than zero. Computed (left) and experimental (right) results are shown for comparison, (a) src SH3 domain; (b) α spectrin SH3 domain; (c) Sso7d. Protein structures are drawn with Molscript.[34]

corresponding results for α-spectrin SH3 domain are plotted in Figure 4(b, left) where we can see that most of the features of the TSE and folding pathways are very similar to those of src SH3 domain, which is also consistent with experimental results shown in Figure 4(b, right). It is interesting to note that in Figure 1(d) the first β-hairpin, called RT loop, gets some native-like order at very early stage of folding, $Q \sim 0.25$. This order has to be broken, however, for further proceed the reaction. By repeated simulation, we confirmed that this is not due to statistical error. The same tendency is found in a free energy functional approach to src SH3 domain[23] and thus quite possibly this is not just an artifact.

Sso7d is a structural analog of the SH3 domain where the fifth β-strand is replaced with an α-helix and it has detectable sequence homology to neither src nor α spectrin SH3 domains. The characteristics of the TSE of Sso7d were experimentally determined[23] and they turned out to be very different from those of src and α-spectrin SH3 domains. Thus, it is an interesting test of the current topology-alone model. Figure 4(c, left) shows computed ϕ-values, which are compared with experimental data given in Figure 4(c, right). We see that the simulated TSE is mostly similar to that of src SH3 domain, and thus deviates significantly from ϕ-value experiment of Sso7d. This suggests that the native topology of this family drives the folding pathways to the TSEs of the src SH3 domains, but the pathways are perturbed by the C-terminal helix and sequence-specific interactions in Sso7d.

U1A, AcP, and ADA2h

Despite low mutual sequence homology, all these proteins have the same super-fold, so-called the αβ plaits composed of βαββαβ secondary structures, and characteristics of the TSEs of them have been anticipated by the protein engineering experiments.[9,10,13] Briefly, the acylphosphatase(AcP) and the activation domain of procarboxypeptidase A2 (ADA2 h) share the characteristic of TSE's, which is an expanded version of the native structure; the central section of the β-sheet (β-strands 1 and 3) is more ordered than the other strands and helices are partially formed.[9,10] On the other hand, the U1A has more polarized TSE in which β strands 2 and 3 with helix 1 (one side of the protein) are organized.[13]

We compare computed ϕ-values with experimental ones in Figures 5. For U1A, computed ϕ-values (Figure 5(a) left) are markedly high at β strands 1 and 3, and helix 1 (left half of Figure 5), which is perfectly consistent with what is anticipated from experiments (Figure 5(a), right). The computed ϕ-values of AcP is plotted in Figure 5(b), left) accompanied by the experimental data in Figure 5(b), right). We see here that the computed TSE significantly deviates from the experimental results and is quite similar to that of U1A. The results for ADA2 h (data not shown) are similar to the case of AcP giving discrepancy from experimental data. In other words, the current topology-only model leads to basically the same TSEs for all of these proteins, implying that the polarized TSE is a direct consequence of the native topology. This suggests two of the three proteins have the central β-sheet formed at their TSEs owing to sequence-specific interactions.

Other cases

The immunoglobin binding domain of protein L and protein G are other interesting examples because they share the same symmetric topology, but the TSEs experimentally observed are different each other: The protein L has the structured N-terminal β-hairpin at the TSE, while the C-terminal hairpin is ordered in the case of protein G. Our simulation gave the high ϕ-values in the N-terminal hairpin for both protein L and protein G (data not shown). Despite a nearly symmetric native topology, the polarized TSE again seems to be a natural consequence of the topology-only model. Nevertheless in the laboratory, the way of breaking the symmetry is not decided by the topology alone, but by sequence-specific information.

We have investigated the characteristics of the TSEs of other proteins, too. Roughly, we find that the computed TSEs are consistent with ϕ-value experiments for about half of the proteins. Proteins with more complex topologies, such as the fibronectin type 3, tend to give better agreement with experimental results than those of simple topology made of α-helices alone, such as λ repressor. For such helical proteins, the TSE characteristic are fragile and can easily be perturbed by changes in parameters.

Discussion

As mentioned above, the scaling law $k_F \sim \exp(-c\ RCO\ N^{0.6})$ found in this paper is slightly different from what was suggested experimentally $k_F \sim \exp(-c\ RCO)$. However, we note that the temperatures in our study are not the same as in the experiments. The experimental data were collected mostly 20-25 °C. This is somewhat lower than T_F. Moreover, since kinetic experiments were performed by many different groups, temperatures used for measurements are different one by one, which makes detailed interpretation difficult. On the other hand, we have performed folding simulations carefully choosing the temperature consistently near T_F leading to the scaling law closer to the theoretical prediction by Finkelstein & Wolynes, $k_F \sim \exp(-c\ N^{2/3})$. This means that the free energy barrier for folding scales as $N^{2/3}$, which is the size of the boundary surface of the folding nucleus at the transition temperature†. The original scaling of their theory itself is, however, hardly detectable at least for small-to-medium size proteins as we see in Figure 2(a) probably because topology effect is so strong that it washes out signal of chain length dependence. Thus, combination of two observations, topology effect and chain-length scaling, may gives a unified rule for folding kinetics near the folding transition temperature.

Wolynes[17] predicted that the chain-length scaling varies from $k_F \sim \exp(-c\ N^{2/3})$ near T_F to a weaker dependence $k_F \sim N^{-\nu}$ at $T \ll T_F$ (but above the glass transition temperature). Thus, probably, as decreasing temperature, the chain-length dependence of the current model becomes weaker, with the scaling of k_F approaching that observed experimentally. Further analysis is needed to this

† For a large protein of N residue long, the gyration radius of the compact form scales as $N^{1/3}$, and global boundary surface area scales as $N^{2/3}$.

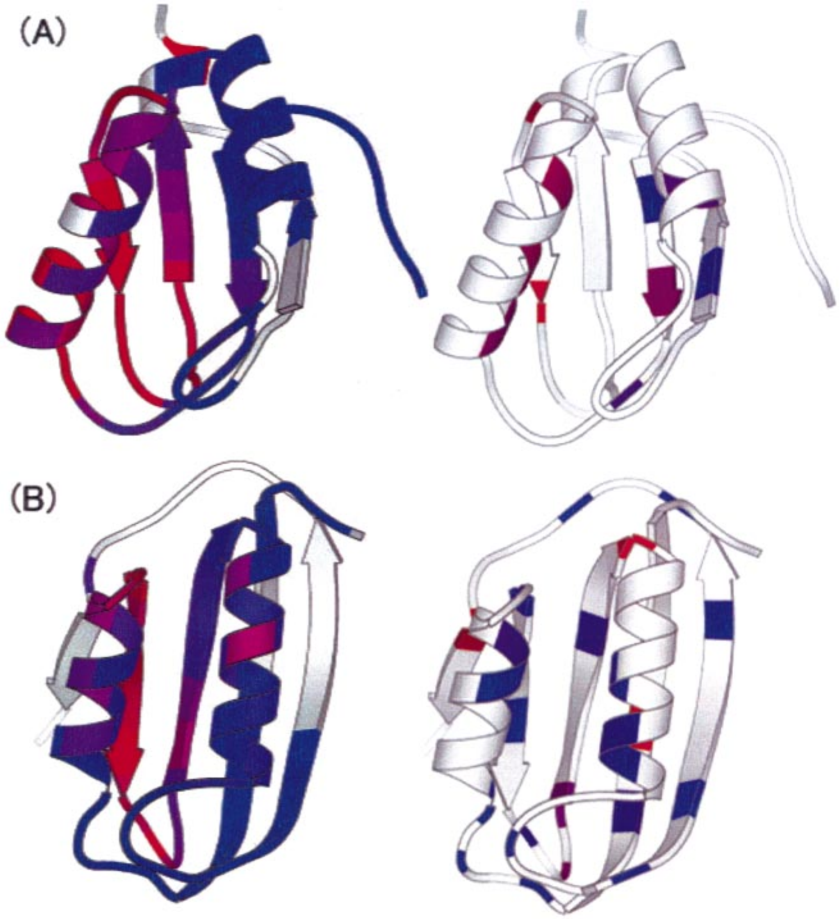

Figure 5. Both experimental and computed ϕ-values of U1A and Acylphosphatase (AcP). Computed (left) and experimental (right) results are shown for comparison; (a) U1A and (b) AcP. Protein structures are drawn with Molscript.[34]

direction. In this context, Cieplak & Hoang[31] found the scaling law of folding rate $k_F \sim N^{-\nu}$ where ν is fitted to around 2.5 by performing folding simulation of a simple protein model somewhat similar to this work. They did the simulation at the temperature where the folding is the fastest. In the current case, the fastest folding is attained at $T \ll T_F$ and thus their observation is consistent with ours based on Wolynes' theory.

For about half of proteins studied, the calculated TSE's are not similar to what was anticipated from experimental ϕ-value analysis. Many reasons can be considered for these disagreements. Comparing with successful prediction by Gō-like free energy functional approaches,[20–23] we consider that a major property to be taken into account would be an inherent propensity of the secondary structure of each segment. The hydrogen bond interaction might be another source of missing specificity. Simulation study with these interactions is under consideration. For some proteins, non-native interactions may play a crucial role in the TSEs, although their importance in general is not clear at the moment.

The straightforward simulation study presented here provides a good tool to validate approximations made in free energy functional approaches. In particular, the contiguous sequence approximation used by Alm & Baker[21] and Munoz & Eaton[22] restricts the number of contiguous native-like segments less than one or two. The validity of this approximation depends on the chain stiffness.[32] With the stiffness used in the current study, at most two contiguous native-like segments seem to be sufficient to describe the folding pathways as we can see in Figure 1(d).

Using the model proposed by Onuchic *et al.*[2] of proteins, we have performed comprehensive folding simulation on 18 single-domain proteins and investigated folding rates and pathways with comparison to experimental results. First, we find that the folding rate constants scales as $k_F \sim \exp(c\ RCO\ N^{0.6})$ where *RCO* and *N* are the relative contact order and number of residues, respectively. This suggests the scaling law of the folding rates $k_F \sim \exp(c\ RCO\ N^{2/3})$ fills the gap between the experimentally observed correlation with the *RCO* and the chain-length scaling property predicted by Finkelstein[15] and Wolynes.[17] Second, the residual native-like order in the denatured state is highly correlated with the *RCO* with negative correlation coefficient; proteins with smaller *RCO* tend to have higher degree of residual order in the denatured states. This is consistent with the observation that many helical proteins possess partial helices in the denatured states. Finally, the topology-only model used in this paper successfully describes the folding pathways and the TSEs for about half of proteins studied.

Materials and Methods

Proteins studied

The proteins swtudied are cytochrome b562 (256b), apo-myoglobin (1pmb), monomeric λ repressor (1lmb), albumin binding domain (1prb), Sso7d (1bnz), Im9 (1imq), ACBD (2abd), cold shock protein B (1nmg), chymotripsin inhibitor 2 (1coa), TN fibronectin type 3 (1ten), U1A (1urn), IgG-binding domain of protein G (2gb1), Ada2 h (1aye), FKBP (1fkt), IgG-binding domain of protein L (2ptl), α-spectrin SH3 domain (1shg), src SH3 domain (1srl) and Acylphosphatase (2acy), where the PDB code is given in parentheses.

Simulation algorithms

The interaction energy E at a given protein conformation Γ is given as

$$\begin{aligned} E(\Gamma, \Gamma_0) = & \sum_{\text{bonds}} K_r (r_i - r_{0i})^2 \\ & + \sum_{\text{angles}} K_\theta (\theta_i - \theta_{0i})^2 \\ & + \sum_{\text{dihedral}} \{ K_\phi^{(1)} [1 - \cos(\phi_i - \phi_{0i}^{(1)})] \\ & + K_\phi^{(3)} [1 - \cos 3(\phi_i - \phi_{0i}^{(3)})] \} \\ & + \sum_{i>j-3}^{\substack{\text{native} \\ \text{contact}}} \varepsilon_1 \left[5 \left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{0ij}}{r_{ij}} \right)^{10} \right] \\ & + \sum_{i>j-3}^{\substack{\text{non-native} \\ \text{contact}}} \varepsilon_2 \left(\frac{C}{r_{ij}} \right)^{12} \end{aligned}$$

In the equation, r_i, θ_i, and ϕ_i stand for the ith virtual bond length between ith and $(i+1)$th amino acid residues, the virtual bond angle between $(i-1)$th and ith bonds, and the virtual dihedral angle around the ith bond, respectively. r_{ij} is the distance between ith and jth amino acid residues. All the parameters with subscript "0" mean the values of the corresponding variables at the native structure. The first term keeps the chain connectivity, while the second and the third terms represent the local torsional interactions. The fourth and fifth terms are non-local interactions, where the former includes native contact interactions and the latter is nonspecific repulsion between non-native pairs. For other parameters, we use $K_\mathrm{r} = 100.0$, $K_\theta = 20.0$, $K_\phi^{(1)} = 1.0$, $K_\phi^{(3)} = 0.5$, $\varepsilon_1 = 0.18$, $\varepsilon_2 = 0.18$, $C = 4.0$ Å for all proteins studied.

We define that the ith and jth amino acid residues are in the native contacts if one of the non-hydrogen atoms in the ith amino acid residue are within a critical distance σ to any non-hydrogen atom in the jth amino acid residue. We tested both $\sigma = 5.5$ Å and 6.5 Å in investigating folding reaction of ten different proteins finding no qualitative difference. Here, we present results for $\sigma = 6.5$ Å.

A standard algorithm[33] is used for performing constant temperature molecular dynamics simulation, where the Newtonian equation of motion is numerically solved with velocity rescaled for keeping the temperature.

The measure of the native-ness $Q(\Gamma)$ is defined, for a given conformation Γ, as the number of formed native contacts divided by the total number of native contacts, where a contact between i and j is classified as formed when r_{ij} is shorter than $1.2 r_{0ij}$.

We estimate the first passage time $t_1 = \langle t_\mathrm{fold} \rangle$ where t_fold is the time when the protein first reaches the native topology, which is judged by the Q-value, and the average $\langle \cdots \rangle$ is taken over 50 trajectories. The folding rate constant k_F is obtained by $k_\mathrm{F} \sim 1/t_1$. The TSEs are obtained from the same 50 trajectories.

Acknowledgments

We thank George Chikenji and Rie Tatsumi for helpful discussions. The work was supported, in part, by Grant-in-Aid for scientific research on priority areas (C) "Genome Information Science" and priority areas (A) "Molecular Physical Chemistry" from the Ministry of Education, Science, Sports and Culture of Japan.

References

1. Fersht, A. R. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, Freeman, New York.
2. Onuchic, J. N., Luthey-Schulten, Z. A. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545-600.
3. Grantcharova, V., Alm, E. J., Baker, D. & Horwich, A. L. (2001). Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* **11**, 70-82.
4. Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, **405**, 39-42.
5. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.
6. Itzhaki, L. A., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol,* **254**, 260-288.
7. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010-1016.
8. Riddle, D. S., Grantcharova, V. A., Santiago, J. D., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016-1024.
9. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005-1009.
10. Villegas, V., Martinez, J. C., Aviles, F. X. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.

11. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971-984.
12. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of β-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669-673.
13. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snapshot to movie: ϕ analysis of protein folding transition states taken one step further. *Proc. Natl Acad. Sci. USA,* **96**, 14854-14859.
14. Gutin, A. M., Abkevich, V. & Shakhnovich, E. I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Letters,* **77**, 5433-5436.
15. Finkelstein, A. V. & Badredtinov, A. Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* **2**, 115-121.
16. Thirumalai, D. (1995). From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys. (France),* **15**, 1457-1467.
17. Wolynes, P. G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl Acad. Sci. USA,* **94**, 6170-6175.
18. Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1997). Structural correlations in protein folding funnels. *Proc. Natl Acad. Sci. USA,* **94**, 777-782.
19. Portman, J. J., Takada, S. & Wolynes, P. G. (1998). Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Letters,* **81**, 5237-5240.
20. Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA,* **96**, 11299-11304.
21. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA,* **96**, 11305-11310.
22. Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **96**, 11311-11316.
23. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967-982.
24. Gō, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183-210.
25. Takada, S. (1999). Go-ing for the prediction of protein folding mechanisms. *Proc. Natl Acad. Sci. USA,* **96**, 11698-11700.
26. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and ''en-route'' intermediates for protein folding? *J. Mol. Biol.* **298**, 937-953.
27. Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000). How native-state topology affects the folding of dihydrofolate reductase and interleukin-1 β. *Proc. Natl Acad. Sci. USA,* **97**, 5871-5876.
28. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. (1992). Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **13**, 1011-1021.
29. Eliezer, D., Yao, J., Dyson, H. J. & Wright, P. E. (1998). Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nature Struct. Biol.* **5**, 148-155.
30. Bai, Y., Karimi, A., Dyson, H. J. & Wright, P. E. (1997). Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.* **6**, 14491457.
31. Cieplak, M. & Hoang, T. X. (2000). Scaling of folding properties in Gō models of proteins. *J. Biol. Phys.* **26**, 273-294.
32. Portman, J. J., Takada, S. & Wolynes, P. G. (2001). Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach. *J. Chem. Phys.* **114**, 5069-5081.
33. Berendsen, H. J., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690.
34. Kraulis, P. J. (1991). A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallog.* **24**, 946-950.

Edited by B. Honig

(Received 20 June 2001; received in revised form 15 August 2001; accepted 15 August 2001)