Discrete molecular dynamics studies of the folding of a protein-like model

Nikolay V Dokholyan[1], Sergey V Buldyrev[1], H Eugene Stanley[1] and Eugene I Shakhnovich[2]

Background: Many attempts have been made to resolve in time the folding of model proteins in computer simulations. Different computational approaches have emerged. Some of these approaches suffer from insensitivity to the geometrical properties of the proteins (lattice models), whereas others are computationally heavy (traditional molecular dynamics).

Results: We used the recently proposed approach of Zhou and Karplus to study the folding of a protein model based on the discrete time molecular dynamics algorithm. We show that this algorithm resolves with respect to time the folding $\rightleftharpoons$ unfolding transition. In addition, we demonstrate the ability to study the core of the model protein.

Conclusions: The algorithm along with the model of interresidue interactions can serve as a tool for studying the thermodynamics and kinetics of protein models.

Addresses: [1]Center for Polymer Studies, Physics Department, Boston University, Boston, MA 02215, USA. [2]Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA.

Correspondence: Nikolay V Dokholyan
E-mail: dokh@bu.edu

Key words: Gō model, molecular dynamics, protein folding

Received: **18 May 1998**
Revisions requested: **23 June 1998**
Revisions received: **07 October 1998**
Accepted: **02 December 1998**

Published: **15 December 1998**
http://biomednet.com/elecref/1359027800300577

Folding & Design 15 December 1998, **3**:577–587

© Current Biology Ltd ISSN 1359-0278

Introduction

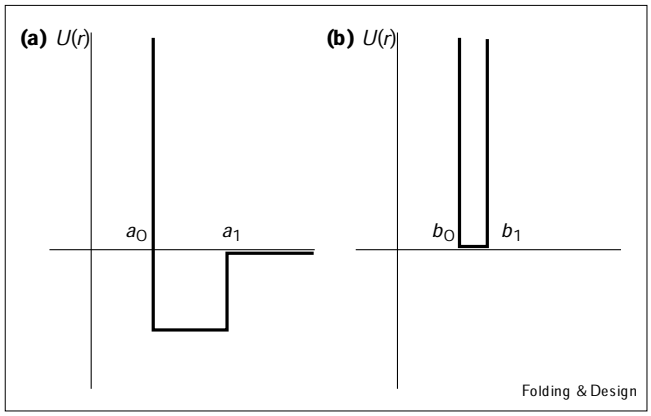
The vast dimensionality of the protein conformational space [1] makes the folding time too long to be reachable by direct computational approaches [2–4]. Simplified models [5–14] became popular due to their ability to reach reasonable time scales and to reproduce the basic thermodynamic and kinetic properties of real proteins [3,15,16]: firstly, a unique native state, that is, there should exist a single conformation with the lowest potential energy; secondly, a cooperative folding transition (resembling first order transition); thirdly, thermodynamic stability of the native state; and fourthly, kinetic accessibility, that is, the native state should be reachable in a biologically reasonable time [12,17].

Monte Carlo (MC) simulations on lattice models (e.g. see [4–7] and references therein) appear to be useful for studying theoretical aspects of protein folding. The MC algorithm is based on a set of rules for the transition from one conformation to another. These transitions are weighted by a transition matrix that reflects the phenomena under study. The simplicity of the algorithm and the significantly small conformational space of the protein models (due to the lattice constraints) make MC on-lattice simulations powerful tools for studying the equilibrium dynamics of protein models; however, lattice models impose strong constraints on the angles between the covalent bonds, thereby greatly restricting the conformational space of the protein-like model. The additional drawback of this restriction lies in the poor capability of these models to discern the geometrical properties of the proteins. The time in MC algorithms is estimated as the average number of moves (over an ensemble of the folding $\rightleftharpoons$ unfolding transitions) made by a model protein. It was pointed out [18] that MC simulations are equivalent to the solution of the master equation for the dynamics, so there is a relation between physical time and computer time, which is counted as the number of MC steps; however, a number of delicate issues — such as the dependence of the dynamics on the set of allowed MC moves — remain outstanding, so an independent test of the dynamics using the molecular dynamics (MD) approach is needed.

To address aspects that are sensitive to geometrical details, it is useful to study off-lattice models of protein folding. Thus far, several off-lattice simulations have been performed [19–21] that demonstrate the ability of the simplified models to study protein folding.

Here, we study the three-dimensional molecular dynamics of a simplified model of proteins [6,7]. The potential of interactions between pairs of residues is modeled by a 'square-well', which allows us to increase the speed of the simulations [22,23]. We estimate folding time based on the collision event list, which, besides increasing the speed of the simulation, allows for the tracking of 'realistic' (not discretized) time. We show that such an algorithm

Figure 1



The potential of interaction between **(a)** specific residues and **(b)** neighboring residues (covalent bond). a_0 is the diameter of the hard sphere and a_1 is the diameter of the attractive sphere. $[b_0 b_1]$ is the interval in which residues that are neighbors on the chain can move freely.

can be a useful compromise between computationally heavy traditional MD and fast but restrictive MC. We demonstrate that the model protein reproduces the principal features of folding phenomena described above.

We also address the issue of whether we can study the equilibrium properties of the core. The core is a small subset of the residues that maintains the backbone of the structure at temperatures close to the folding transition temperature (here the Θ-temperature, T_θ). We emphasize the difference between the core and the nucleus of a protein: whereas the core is a persistent part of the structure at equilibrium, the nucleus is a fragment of this structure that is assembled in the transition state (TS) — the folding $\rightleftharpoons$ unfolding barrier (see Figure 1 in [4]). Based on simple arguments, we estimate T_θ for our model [24] and compare it with the value found in the simulations.

The model

We study a 'beads-on-a-string' model of a protein. We model the residues as hard spheres of unit mass. The potential of interaction between residues is 'square-well'. We follow the Gō model [5–7], where the attractive potential between residues is assigned to the pairs that are in contact (Δ_{ij}, defined below) in the native state and repulsive potential is assigned to the pairs that are not in contact in the native state. Thus, the potential energy is:

$$E = \frac{1}{2}\sum_{i,j=1}^{N} U_{i,j} \qquad (1)$$

where i and j denote residues i and j. $U_{i,j}$ is the matrix of pair interactions:

$$U_{i,j} = \begin{cases} +\infty, & |r_i - r_j| \le a_0 \\ -\text{sign}(\Delta_{ij})\epsilon, & a_0 < |r_i - r_j| \le a_1 \\ 0, & |r_i - r_j| > a_1 \end{cases} \qquad (2)$$

Here $a_0/2$ is a radius of the hard sphere, and $a_1/2$ is the radius of the attractive sphere (Figure 1a) and ϵ sets the energy scale. $||\Delta||$ is a matrix of contacts with elements:

$$\Delta_{ij} \equiv \begin{cases} 1, & |r_i^{NS} - r_j^{NS}| \le a_1 \\ -1, & |r_i^{NS} - r_j^{NS}| > a_1 \end{cases} \qquad (3)$$

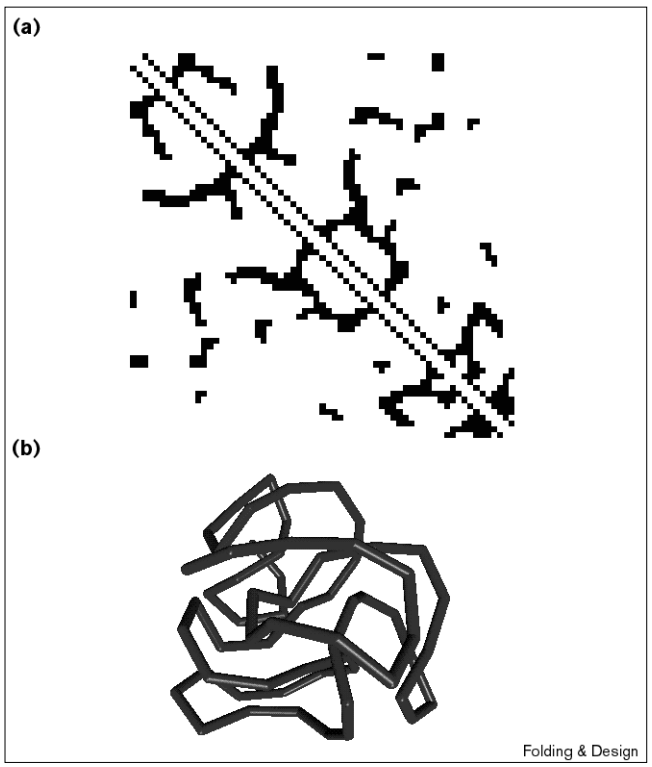
where r_i^{NS} is the position of the i^{th} residue when the protein is in the native conformation. Note that we penalize the non-native contacts by imposing $\Delta_{ij} < 0$. The parameters are chosen as follows: $\epsilon = 1$, $a_0 = 9.8$ and $a_1 = 19.5$. The covalent bonds are also modeled by a square-well potential (Bellemans' bonds):

$$V_{i,i+1} = \begin{cases} 0, & b_0 < |r_i - r_{i+1}| < b_1 \\ +\infty, & |r_i - r_{i+1}| \le b_0, \text{ or } |r_i - r_{i+1}| \ge b_1 \end{cases} \qquad (4)$$

The values of $b_0 = 9.9$ and $b_1 = 10.1$ are chosen so that average covalent bond length is equal to 10 (Figure 1b). The original configuration of the protein ($N = 65$ residues) was designed by collapse of a homopolymer at low temperature [20,25,26]. It contains $n^* = 328$ native contacts, so $E_{NS} = -328$. The 65×65 matrix of contacts of the globule in the native state is shown in Figure 2a. Note that the large number of native contacts ($328/65 \approx 5$ contacts per residue) is due to the choice of the parameter: $a_1 \approx 2a_0$ — so that residues are able to establish contacts with the residues in the second neighboring shell. The radius of gyration of the globule in the native state is $R_G \approx 22.7$. A snapshot of the globule in the native state is shown in Figure 2b.

The program employs the discrete MD algorithm, which is based on the collision list and is similar to one recently used by Zhou *et al.* [22] to study equilibrium thermodynamics of homopolymers and by Zhou and Karplus [23] to study equilibrium thermodynamics of folding of a model of *Staphylococcus aureus* protein A. A detailed description of the algorithm can be found in [27–30]. To control the temperature of the protein, we introduce 935 particles that do not interact with the protein or with each other in any way but via regular collisions, serving as a heat bath. Thus, by changing the kinetic energy of those 'ghost' particles, we are able to control the temperature of the environment. The ghost particles are hard spheres of the same radii as the chain residues and have unit mass. Temperature is measured in units of ϵ/k_B. The time unit (tu) is estimated

Figure 2



(a) 65 × 65 contact matrix of the model protein in the native state. Black boxes indicate the matrix elements of those residue pairs that have a contact (their relative distance is between a_0 and a_1). **(b)** A snapshot of the protein of 65 residues in the native state obtained at temperature $T = 0.1$.

from the shortest time between two consequent collisions in the system between any two particles.

Results

In order to study the thermodynamics, we performed MD simulations of the chain at various temperatures. We start with the globule in the native state at temperature $T = 0.1$ and then raise the temperature of the heat bath to the desired one. Then we allow the system to equilibrate. At the final temperature, we let the protein relax for 10^6 time units. The typical behavior of the energy E and the radius of gyration R_G as functions of time is shown in Figure 3 for three different temperatures.

In the present model, the non-native contacts (NNCs) are penalized, that is, the pairwise interaction between NNCs is repulsive (this corresponds to $g = 2$ in [23]), so their number increases as the temperature increases. At high temperatures (above T_θ), however, the number of NNCs varies only due to the random temperatures. The maximum number of NNCs occurs at T_θ and does not exceed 35, which is roughly 10% of the total number of native contacts (NCs).

The simulations reveal that the protein undergoes a folding $\rightleftharpoons$ unfolding transition as we increase the temperature to the proximity of the Θ-temperature T_θ, which in this model is $T_\theta \equiv T_f \approx 1.46$. At T_θ, the distribution of energy has three peaks (Figure 4a). The left peak corresponds to the folded state, the right peak corresponds to the unfolded state and the middle one corresponds to the partially folded state (PFS), with a 19-residue unfolded tail. This trimodality of the energy distribution is also seen in Figure 3b. The energy profile at temperature $T = 1.42$ (close to T_θ) also reflects these three states. Since $T < T_\theta$, only two states are mostly present in Figure 3b. Thus, the energy distribution has only two peaks (Figure 5), corresponding to the folded state and the PFS. Above T_θ, the globule starts to explore energetic wells other than the native well (see Figure 13 in [31]).

To show that the PFS is the cause of the middle peak in energy distribution (Figure 4a), we eliminate the 19-residue tail and plot the energy distribution for the 46-mer at Θ-temperature $T_\theta^* = 1.44$ (Figure 6). We expect to see only two states — folded and unfolded — because the 19-residue tail, which is the cause of the PFS, is eliminated. Figure 6 confirms our expectations.

The folding $\rightleftharpoons$ unfolding transition is further quantified in Figure 7. The energy and the radius of gyration increase most rapidly near $T_f = T_\theta$ resembling the order parameter jump in a phase transition (see discussion below). This rapid increase of E and R_g reaches its maximum at the Θ-point, where the potential of interaction is compensated by the thermal motion of the particles. Above T_θ, interactions between residues do not hold them together any more and the chain becomes unfolded (see Figure 8a). Note that as all the attractive interactions are specific, the transition is described by one temperature, T_f.

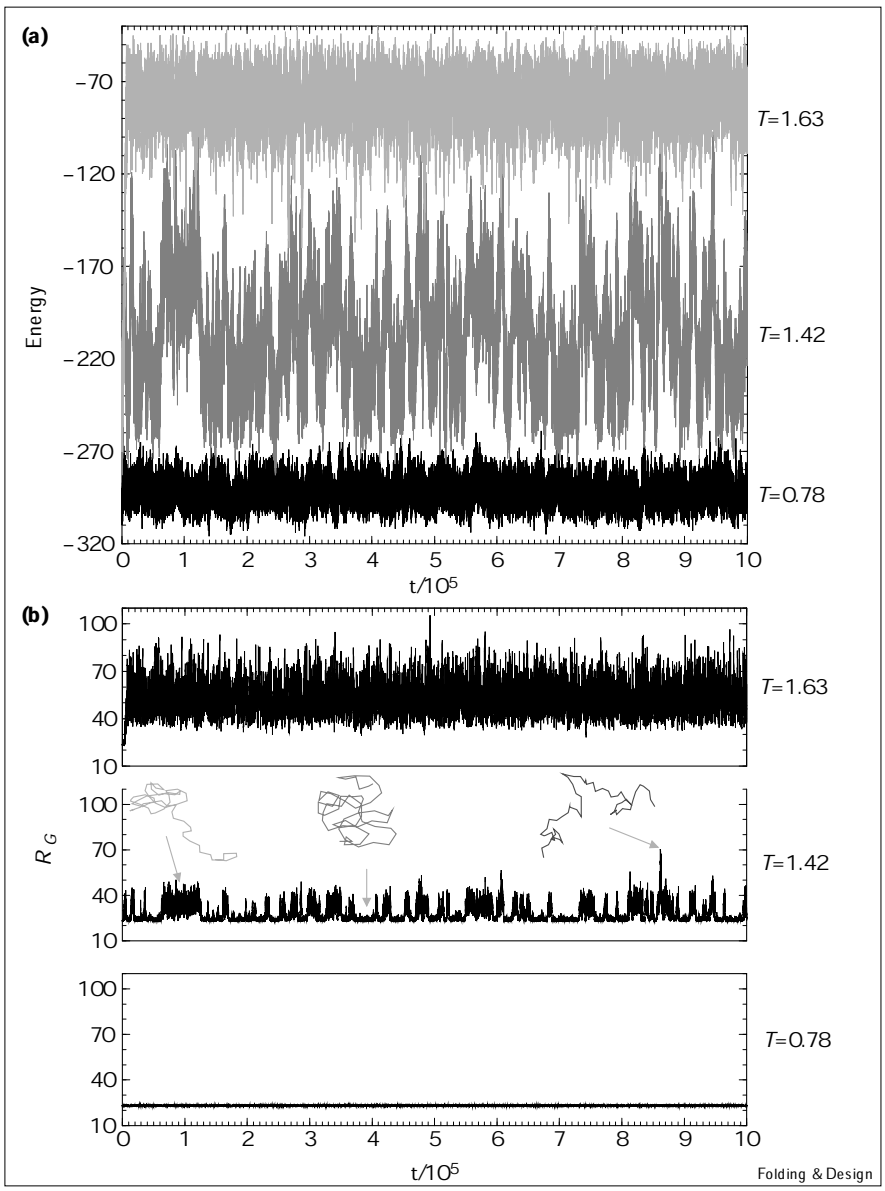
The presence of the PFS is observed in the temperature range, 1.40 to 1.48, in which the collapse transition occurs. Thus, in this particular region, T_f and T_θ are indistinguishable within the accuracy of their definitions.

Remarkably, a simple Flory-type model of an excluded volume chain predicts T_θ within 20%. To demonstrate this, let us write the probability that the end-to-end type distance of the chain is R [24]:

$$P(R) \propto p(R) \exp\left(-\frac{N^2 v}{2R^3} - \frac{\mathrm{E}(R)}{T}\right) \qquad (5)$$

where $v = (4\pi/3(a_0/2)^3$ is the volume of the monomer and $p(R) \propto R^2 \exp(-3R^2/(2N(a_0/2)^2))$ is the probability that the end-to-end distance of the chain is R for the random walk model. For $T = T_f$, the repulsive excluded volume term $-(N^2 v)/(2R^3)$ balances the attractive term $-\mathrm{E}(R)/T > 0$. Thus:

Figure 3



The dependence on time of **(a)** energy E and **(b)** radius of gyration R_G. The globule is maintained at three different temperatures $T = 0.78 < T_f$, $T = 1.42$, and $T = 1.63 > T_f$ for 10^6 tu. For $T = 0.78$, the fluctuations of both energy E and R_G are small, that is, the globule is found in one folded configuration. At high temperatures ($T = 1.63$) the fluctuations of E and R_G are large; the globule is mostly found in the unfolded state. At the temperature $T = 1.42$, which is close to T_f, the globule is mostly present in two states. The lower energy configuration corresponds to the folded state: the globule is compact —see (b). The other configuration has large fluctuations: the globule is in the PFS. There is an additional state: the unfolded state —see (b). At $T = 1.42$ the protein model is rarely present in the unfolded state. Thus, the behavior of the globule at temperatures close to T_f indicates the presence of three distinct states: folded, unfolded and PFS.

$$T_f = \frac{2R^3|\mathrm{E}|}{N^2 v} \approx 1.7 \qquad (6)$$

where $E \simeq -130$ and $R \simeq 24$ are taken for a certain configuration at the Θ-point.

We also compute the heat capacity C_V from the relation [32]:

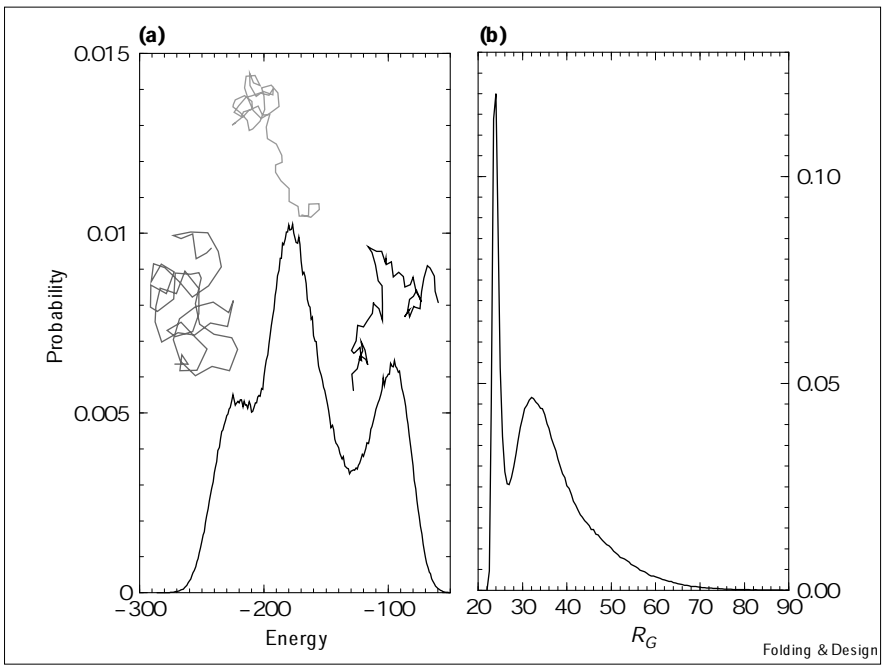
$$C_V = \frac{\left\langle (\delta \mathrm{E})^2 \right\rangle}{T^2} \qquad (7)$$

where $\langle(\delta \mathrm{E})^2\rangle \equiv \langle \mathrm{E}^2\rangle - \langle \mathrm{E}\rangle^2$ and $\langle ... \rangle$ denotes a time average. The time average is computed over 10^6 tu of equilibration at a fixed temperature. The dependence of the heat capacity on temperature is shown in Figure 7b. There is a pronounced peak of $C_V(T)$ for $T = T_f$.

We note that below the folding temperature T_f, the globule (Figure 8b) spends time in a state structurally similar to the native state (Figure 2b); however, one can see that even though the globule maintains approximately the same structure, that is, the same set of NCs, the distances between residues are much larger than in the native state. Due to the fact that the potential of interaction between like residues is a square-well, there is no penalty for these residues to be maximally separated, yet they remain within the range of attractive

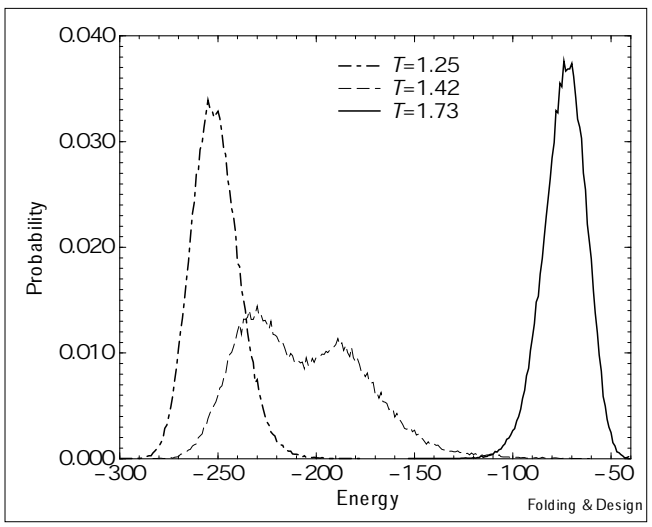
Figure 4

The probability distribution of **(a)** the energy states E and **(b)** the radius of gyration R_G of the globule maintained at $T_f = 1.46$ for 10^6 tu. The trimodal distributions indicate the presence of three states: the folded state, the PFS, and the unfolded state.



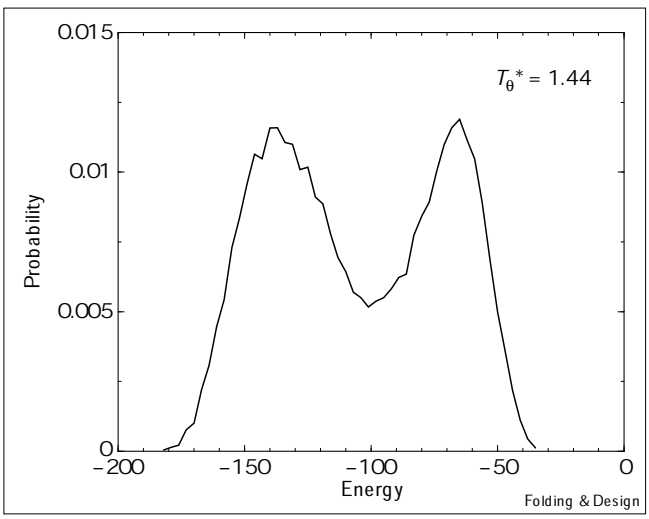
interaction. This allows the globule to have more NNC and, thus, still maintain its similarity to native structure, yet to have energy larger than the energy of the native state. This structure can be identified as the highest in energy that still maintains its core. As the temperature increases, the ratio $|R_G - R_G^{NS}|/R_G^{NS}$ increases until the temperature reaches $T = T_f$, where the ratio becomes roughly 0.87.

Figure 5



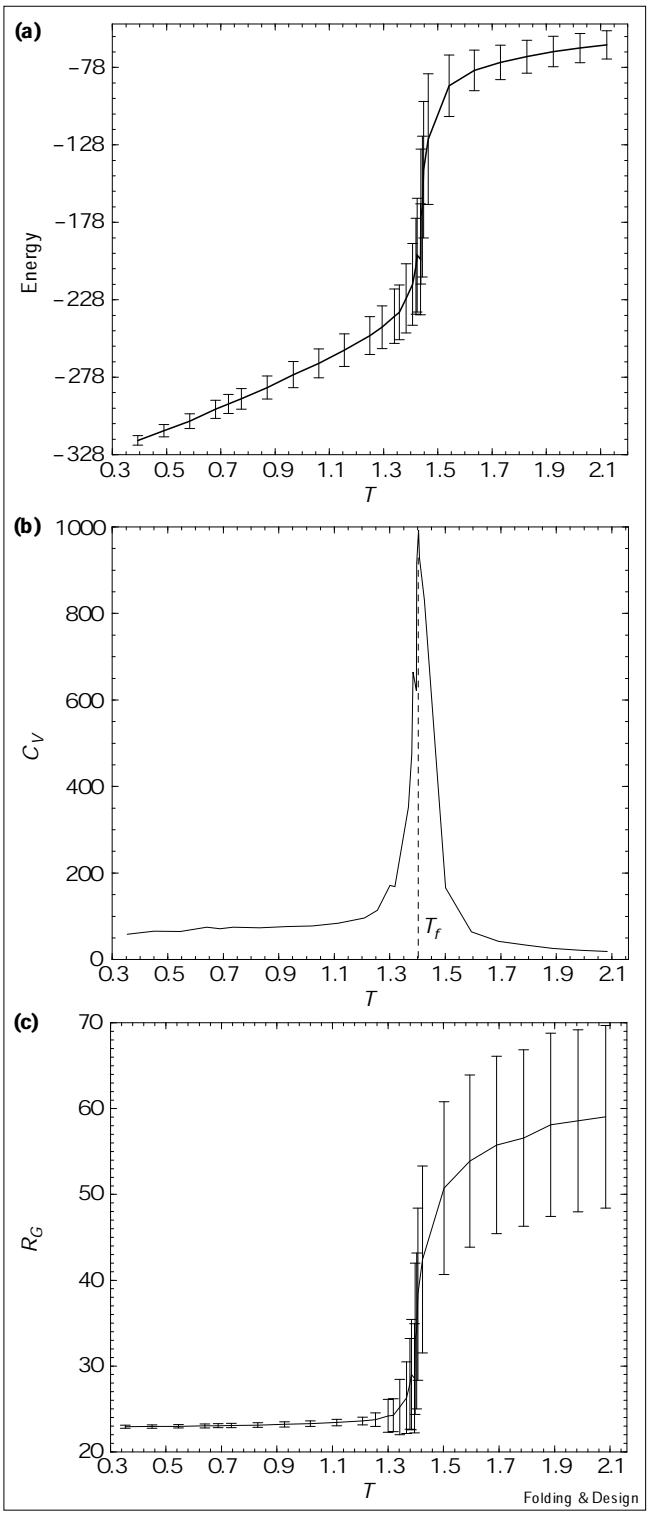
The probability distribution of the energy states E of the globule maintained at three different temperatures: $T = 1.25$, 1.42 and 1.73. Note that at $T = 1.42 \approx T_f$ the distribution has two expressed peaks. The right peak of this ($T = 1.42$) distribution corresponds to the PFS whereas the left peak corresponds to the energetic well of the native state.

Figure 6



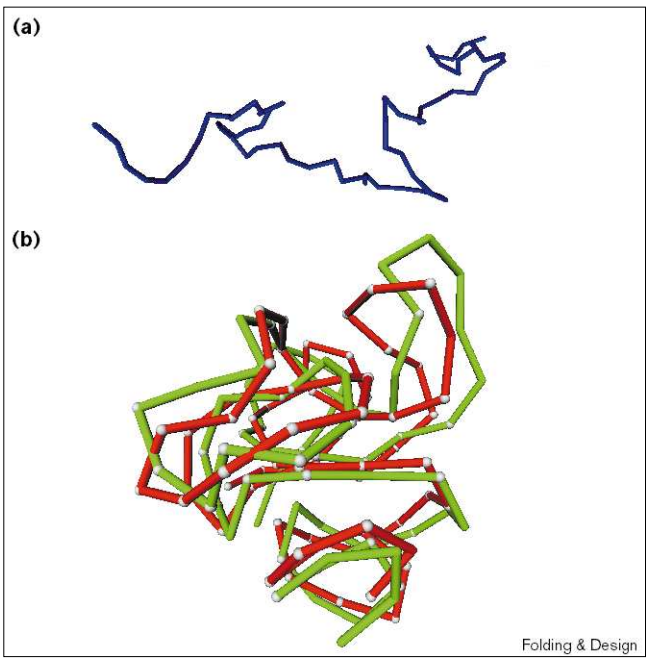
The probability distribution of the energy states E of the 46-residue globule maintained at $T_f^* = 1.44$ during 10^6 tu. The bimodal distributions of energy indicate that the 19-residue tail is responsible for the PFS of the 65-residue globule: after eliminating the 19-residue tail, the trimodal energy distribution of the 65-residue globule becomes a bimodal energy distribution of the 46-residue globule.

Figure 7



The dependence on temperature of **(a)** the energy E, **(b)** the heat capacity C_V and **(c)** the radius of gyration R_G. The error bars are the standard deviation of fluctuations. The rapid increase of energy as well as the sharp peak in heat capacity at $T = T_f$ indicates a first-order phase transition.

Figure 8



A snapshot of the protein in **(a)** the unfolded state, obtained at high temperature $T = 1.8$; and **(b)** the transition state, obtained at folding transition temperature $T_f = 1.46$ (green), overlapped with the globule at low temperature $T = 0.4$ (red). Note that the TS globule has a close visual similarity to those maintained at low temperature and in the native state (see also Figure 2b). It is more dispersed, however, which makes all the NCs easily breakable. To compare the globule at the TS with the one maintained at temperature $T = 0.4$, we perform the transformation proposed by Kabsch [41] to minimize the relative distance between the residues in the TS and the state at $T = 0.4$. The 'cold' residues (grey spheres) denote residues whose rms displacement is smaller than a_1.

To confirm the presence of the core, we calculate $f \equiv N_{NC}/N_C$ at temperatures below $T = T_f$. The attractive interresidue interaction term $-\mathrm{E}/T$ dominates the excluded volume repulsion term $-N^2v/(2R^3)$ (see Equation 5), so:

$$-\frac{\mathrm{E}}{T_f} - \frac{N^2 v}{2R^3} > 0 \qquad (8)$$

The total energy E has contributions from both NC and NNC contacts, so:

$$\mathrm{E} = -\epsilon(N_{NC} - N_{NNC}) = -[2f - 1]\epsilon N_C \qquad (9)$$

At a temperature slightly below T_f, $|T - T_f|\,T_f \approx 0.3$, the residues are maximally separated within their potential wells, yet they still maintain contacts. Therefore, the volume $\tilde{v}$ spanned by one residue is roughly $\tilde{v} \approx (4\pi/3)(a_1/2)^3 = 8v$. N_C is the product of the probability $\tilde{v}/R^3$ of having a bond (NC or NNC) and the total number

of possible arrangements of the pair contacts between N residues, $N(N-1)/2 \approx N^2/2$. Thus:

$$N_C = \frac{N^2}{2R^3}\tilde{v} \tag{10}$$

From Equations 8–10 we can estimate f, the fraction of N_C at the temperature $T \approx 1.42 < T_f$:

$$f > \frac{1}{2} + \frac{v}{\tilde{v}}\frac{T}{\epsilon} \approx 0.68 \tag{11}$$

Due to the fact that the globule maintains roughly the same volume at temperatures slightly below Θ-point, Equation 11 implies that approximately 70% of all native contacts stay intact in the folded phase (see Figure 5). This result is supported by the simulations: at $T \approx 1.42$ the number of NNCs is roughly $N_{NNC} \approx 28$, and the energy E is E = –206. Therefore, the number of NCs is $N_{NC} \approx 234$, and the fraction of NCs is $f \approx 0.89$, which is even higher than the lower limit set by Equation 11. Note that at a temperature higher than T_f, the fraction of native contacts becomes small due to the fact that in this regime the interactions are dominated by the excluded volume repulsion. This change in the number of NCs from 70% to close to zero indicates the presence of the core structure maintained by these 70% of NCs (see Figure 8b and discussion below). Above the Θ-point, the globule is completely unfolded (Figure 8a).

The formation of a specific nucleus during the folding transition was suggested by many theoretical [2,4,11,13,33–36] and experimental works [37–40]. The presence of the core at T_f may arise from a nucleation process driving the system from the unfolded state to the native state. We find indications of a first order transition. We also offer theoretical reasoning for the presence of a core (Equation 11) that might indicate the presence of a nucleus. Next, we identify the core.

We calculate the mean square displacement $\sigma(T)$ of the globule at a certain temperature from a globule at the native state, that is:

$$\sigma(T) \equiv \left\langle \left[\frac{1}{N}\sum_{i=1}^{N} \left(\vec{r}_i^{NS} - \hat{R}\left(\hat{\tau}\vec{r}_i\right) \right)^2 \right]^{1/2} \right\rangle = \left\langle \left[\frac{1}{N}\sum_{i=1}^{N} \sigma_i^2(T) \right]^{1/2} \right\rangle \tag{12}$$

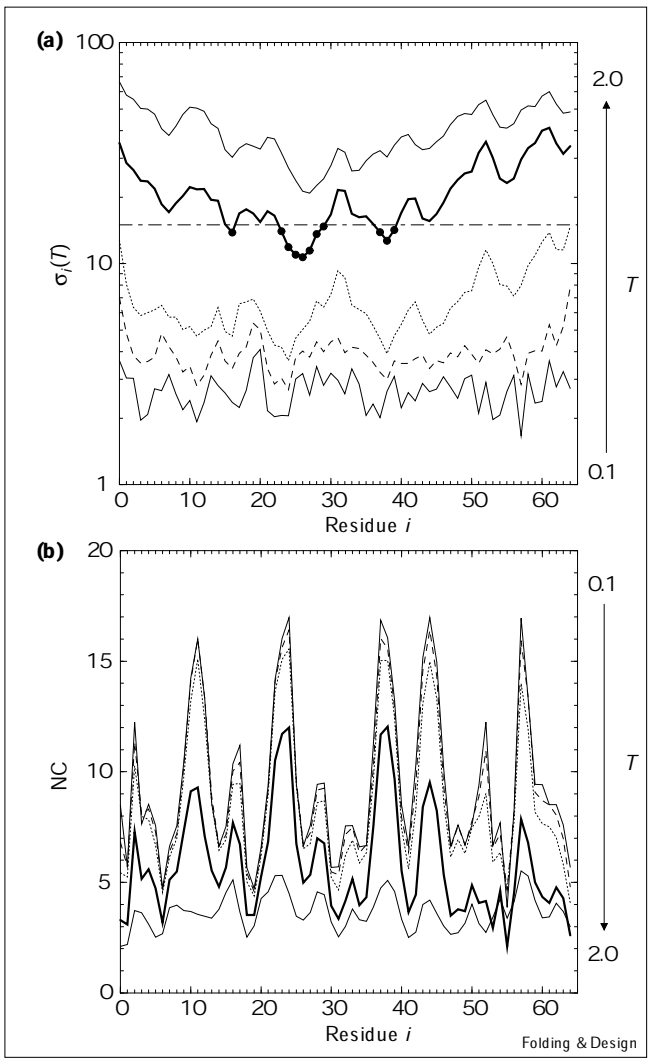
where $\vec{r}_i$ and $\vec{r}_i^{NS}$ are the coordinates of the residues of the globules at two conformations: at some conformation at the temperature T and at the native conformation, respectively. $\hat{T}$ is a translation matrix, which sets the centers of mass of these configurations at the same point in space. $\hat{R}$ is a rotation matrix that minimizes the relative distance between the residues of two configurations (for details, see [41–44]). The $\sigma_i(T)$ values in Equation 12 are the root-mean-square (rms) displacements for each individual residue.

The plot of $\langle\sigma_i(T)\rangle$ is presented in Figure 9a — from the roughness of the 'landscape', we can select a group of residues whose rms displacements are significantly smaller than the rms displacements of the other group of residues. We denote the former group by 'cold' residues and the latter group by 'hot' residues. The rms displacement strongly depends on the temperature near the folding transition and grows slowly below T_f. Note that the average numbers of NC of the residues correlate with the average rms displacement of these residues, that is, the peak on the $N_{NC,i}$ isothermal lines of Figure 9b correspond to the 'cold' residues.

Next we calculate the rms displacement $\sigma_C(T)$ for the selected 25% coldest residues (the core) and $\sigma_O(T)$ for the rest of the residues. Figure 10 shows their dependence on temperature, as well as the dependence of the rms displacement for all residues $\sigma(T)$. There is a pronounced difference in the behavior of the rms displacement of the core residues and the rest of the residues below T_f. At T_f their behavior is the same, due to the fact that all the attractive interactions are balanced by the repulsion of the excluded volume. Above T_f, the difference between $\sigma_C(T)$ and $\sigma_O(T)$ is due only to the fact that the core residues have most of the NCs and, therefore, are more likely to spend time together even at $T > T_f$.

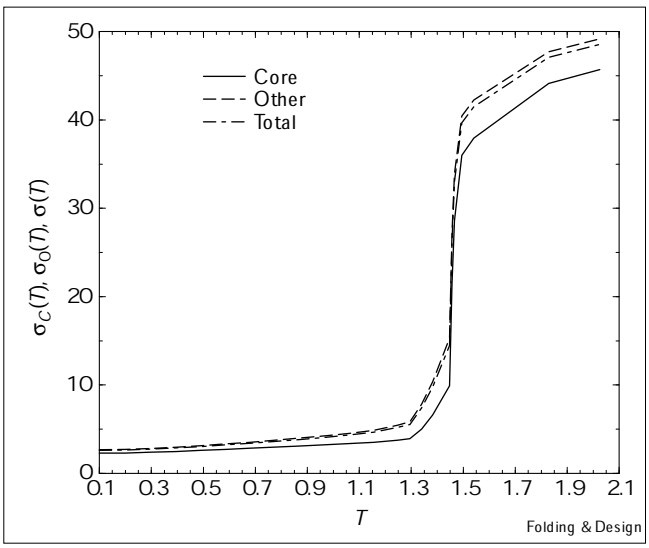
To study the behavior of the globule at T_f, we subdivide the probability distribution of the energy states E of the globule maintained at $T_f = 1.46$ during 10^6 tu into five regions: A, B, C, D, and E (Figure 11a). Region A corresponds to the folded state; region B corresponds to the transitional state between the folded state and the PFS; region C corresponds to the PFS; region D corresponds to the transitional state between the PFS and the completely unfolded state (region E). Next we plot the rms displacement for each residue for each of the above regions (Figure 11a). Note that in region A, all residues stay in contact; in region C, both N- and C-terminal tails break away forming a PFS; in region D, there are only a few core residues that still stay intact and in region E none of the residues is in contact. In region B, we observe that some of the C-terminal tail residues are not in contact, indicating the formation of a PFS. Next, we plot the dependence of the selected 11 core residues (see legend to Figure 9) on the average energy of the window of the corresponding region (Figure 11c). We observe that core residues remain close to one another even in the

Figure 9



(a) The contour plot of the rms displacement $\sigma_i(T)$ for each residue $i = 0, 1, \ldots, 64$ at temperatures $T = 0.3, 0.97, 1.34, 1.46$ (bold line) and 1.54, averaged over 10^6 tu. Note that there is a distinct difference between the 'cold' (small values of $\sigma_i(T)$) and 'hot' (large values of $\sigma_i(T)$) residues. The horizontal line indicates the breaking point of the NCs, that is, when $\sigma_i(T)$ is of the size of the average relative position between pairs of residues, that is, $\sigma_i(T) = (a_0 + a_1)/2 \approx 15$. The bold lines —in both (a) and (b) —indicate the folding transition temperature line T_f. It is worth noting that 11 residues are still in contact —marked by circles on (a): 16, 23, 24, 25, 26, 27, 28, 29, 37, 38 and 39. **(b)** An analogous plot to (a) of the average number of NCs for each residue. Note that the number of NCs correlates strongly with therms: the local minima of the $\langle\sigma_i(T)\rangle$ plots correspond to the local maxima of the number of NCs.

Figure 10



The dependence of the rms displacement of the core residues $\sigma_C(T)$, the rest of the residues $\sigma_O(T)$ and all the residues $\sigma(T)$ on temperature. The above quantities are averaged over 10^6 tu. Note that for the ideal first-order phase transition, one would expect $\sigma_C(T)$ to be a step function; however, as we consider a transition that would be first order in the limit of the infinite size, $\sigma_C(T)$ exhibits only step-function-like behavior. The difference between core residues and other residues is that at T_f the average rms displacement of the core residues is smaller than 15, which indicates that they are in contact (see the legend to Figure 9). On the contrary, the average rms displacement of the non-core residues is greater than 15, indicating that these residues are not in contact.

second transitional state D between the PFS and the completely unfolded state.

We also study the system by cooling it from the high temperature state. This technique corresponds to simulated annealing, due to the fact that the temperature control is governed by the ghost particles that are present in the system. We find that if the target temperature is above 1.1, the globule always reaches the state corresponding to the native state; however, if the target temperature is 0.96, the globule reaches the state corresponding to the native state in only ≈ 70% of cases, in the time interval of 10^5 time units. As an example, we demonstrate in Figure 12 the cooling of the model protein from the high temperature state $T = 3.0$ to the low temperature state $T = 0.1$. The model protein collapses after 1200 tu.

What is particularly remarkable about Figure 12 is that we can follow the kinetics of the collapse. First, the globule gets trapped in some misfolded conformation, where it stays for about 1000 tu (see Figure 12a) and then it collapses to the native state. The time behavior of the energy, however, can look a bit puzzling. After the rms displacement drops to close to 0, indicating the native state, the energy is still higher than that of the native state for about 10^4 tu (see Figure 12b). The key to resolving this puzzle is the fact that after the collapse of the model protein, its potential energy transforms to kinetic energy,

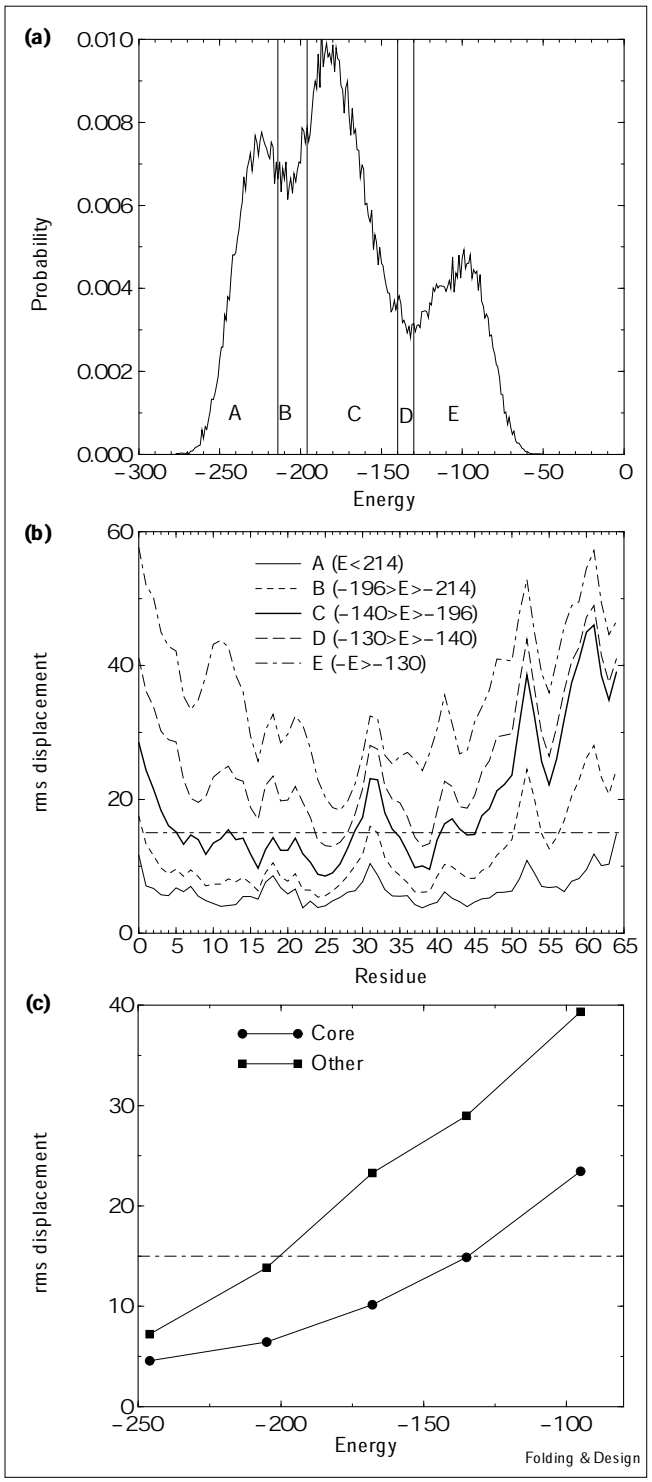


Figure 11

(a) The probability distribution of the energy states E of the globule maintained at $T_f = 1.46$ during 10^6 tu. The probability distribution is divided into five regions: A, B, C, D and E. Region A corresponds to the folded state; region B corresponds to the transitional state between the folded state and the PFS; region C corresponds to the PFS; region D corresponds to the transitional state between the PFS and the completely unfolded state (region E). **(b)** The plot of the rms displacement $\sigma_i(T)$ for each residue $i = 0, 1, \ldots, 64$ for regions A, B, C, D and E from the plot in (a) averaged over 10^6 tu. Note that in region A, all residues stay in contact; in region C, both N- and C-terminal tails break away, forming a PFS; in region D, there are only a few of the core residues that are still intact; and in region E, none of the residues is in contact. **(c)** The dependence of the rms displacement of the core residues (circles) and the other residues (squares) on the average energy E of the window of the corresponding region. Note that core residues stay intact even in the second transitional state D between the PFS and the completely unfolded state. The horizontal lines in (b,c) indicate the breaking point of the NCs (see Figure 9).

which slowly decreases by thermal equilibration with the bath of the ghost particles.

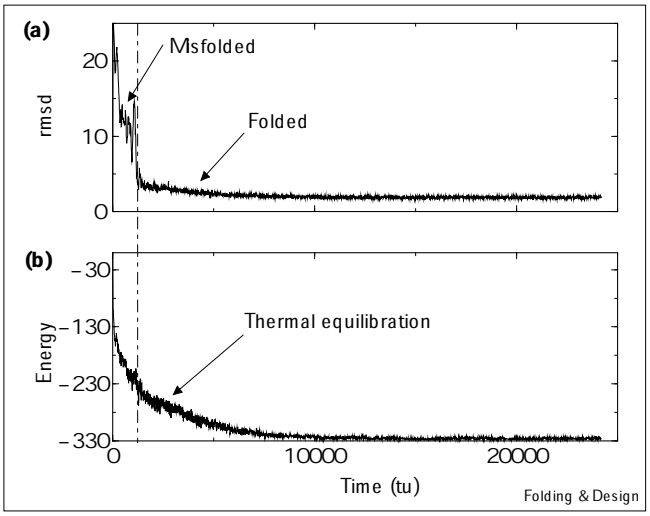
Discussion

We find that the classical model of the self-avoiding chain with excluded volume shows good agreement with the simulations. We show from simple arguments and simulations that the fraction of NCs at the folding temperature T_f is larger then 70%, consistent with the presence of the core. The nucleus forms in the unstable transition state. From the transition state, the globule jumps either to the completely unfolded conformation or to the folded conformation.

Our simulations are in agreement with the recent work of Zhou and Karplus [23]. They performed discrete MD simulations of *S. aureus* protein A, the interresidue interactions of which were modeled based on the Gō model [5–7]. The pair residues of the model protein, which form native contacts, had 'square-well' potential of interaction with the depth of the well equal to $B_N\epsilon$, whereas all other pair residues had 'square-well' potential of interaction with the depth of the well equal to $B_O\epsilon$. They characterized the difference between NCs and NNCs by the 'bias gap', g: $g = 1 - B_O/B_N$. Zhou and Karplus found that when $g = 1.3$, that is, when the interaction between NCs is of the opposite sign to the interaction between NNCs, there is a strong first-order-like transition from the random coil to the ordered globule. The case with our globule corresponds to $g = 2$, where, according to the work of Zhou and Karplus, there should exist a strong first-order-like transition from the random coil to the ordered globule without intermediate.

We also select the core residues and show that their rms displacement behaves significantly differently to the behavior of the rms displacement of the rest of the residues and exhibits step-function-like behavior upon the change of temperature. Our findings are in agreement with the recent experimental study of the equilibrium hydrogen exchange behavior of cytochrome *c* of Bai *et al.* [39], who investigated the exposure of the amide hydrogens (NH) in cytochrome *c* to solvent (due to local and global unfolding fluctuations). The experiments were based on the properties of the amide hydrogens

Figure 12



The time evolution of the **(a)** rms displacement per residue of the globule from its native state and **(b)** energy when we cool the system from the high temperature ($T = 3.0$) unfolded state down to the low temperature ($T = 0.1$) state. The model protein gets trapped in the misfolded conformation after 200 tu and then proceeds to its native state after 1000 tu. Although the rms displacement of the globule from its native state is close to 0 after 1200 tu, the energy of the globule is higher than the energy of the native state for about 10^4 tu. This effect is due to the thermal bath ghost particles that thermally equilibrate the system during $t_{relax} \approx 10^4$ tu relaxation time.

that are involved in hydrogen-bonded structure and can exchange with solvent hydrogens. Bai *et al.* [39] demonstrated that proteins undergo folding $\rightleftharpoons$ unfolding transition "…through intermediate forms". They also selected these intermediate forms (cooperative units), which are 15–25 residues in size. The presence of PFSs in our simulations is thus in agreement with the finding by Bai *et al.* [39] of the intermediate forms in cytochrome *c*.

The relation between core residues that we find and the nucleus is hard to establish due to the fact that the TS is very unstable. Recent amide hydrogen exchange experiments on the CheY protein from *Escherichia coli* of Lacroix *et al.* [40] provided evidence for the residues being involved in the folding nucleus; furthermore, the lattice MC simulations of Abkevich *et al.* [11] also demonstrate that the presence of the nucleus is a necessary and sufficient condition for subsequent rapid folding to the native state. The crucial difference between the nucleus and the rest of the structure is in dynamics, which is manifest also in equilibrium fluctuations. All local unfolding fluctuations (i.e. the ones after which the chain returns rapidly back to the native state) keep the nucleus intact, whereas fluctuations that disrupt the nucleus lead to global unfolding: 'descend' to the 'unfolded' free energy minimum [4,11]. This view is consistent with the hydrogen exchange experiments [39,40]. Such behavior of the globule is consistent with a possible first-order phase transition in a system of finite size.

Acknowledgements

We would like to thank G.F. Berriz for designing the globule, V.I. Abkevich, L. Mirny, M.R. Sadr-Lahijani and S. Erramilli for helpful discussions, and R.S. Dokholyan for help in editing the manuscript. N.V.D. is supported by NIH NRSA molecular biophysics predoctoral traineeship (GM08291-09). E.I.S. is supported by NIH grant RO1-52126.

References

1. Levinthal, C. (1968). Are there pathways for protein foldings? *J. Chim. Phys.* **65**, 44.
2. Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183-210.
3. Karplus, M. & Shakhnovich, E.I. (1994). Protein folding: theoretical studies of thermodynamics and dynamics. In *Protein Folding.* (Creighton, T., ed). pp. 127–196, W.H. Freeman and Company, New York.
4. Shakhnovich, E.I. (1997). Theoretical studies of protein-folding, thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
5. Taketomi, H., Ueda, Y. & Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulations. *Int. J. Peptide Protein Res.* **7**, 445.
6. Go, N, & Abe, H. (1981). Noninteracting local-structure model of folding and unfolding transition in globular proteins. *I. Formulation. Biopolymers* **20**, 991-1011.
7. Abe, H. & Go, N. (1981). Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers* **20**, 1013-1031.
8. Dill, K.A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501-1509.
9. Bryngelson, J.D. & Wolynes, P.G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **93**, 6902-6915.
10. Shakhnovich, E.I. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **72**, 3907-3910.
11. Abkevich, V.I. Gutin, A.M, & Shakhnovich, E.I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
12. Gutin, A.M, Abkevich, V.I. & Shakhnovich, E.I. (1995). Evolution-like selection of fast-folding model proteins. *Proc. Natl Acad. Sci. USA* **92**, 1282-1286.
13. Shakhnovich, E.I. Abkevich, V.I. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature* **379**, 96-98.
14. Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.
15. Creighton, T. (1992). *Protein Structure and Molecular Properties.* Freeman, San Francisco.
16. Privaov, P.L. (1989). Thermodynamic problems of protein structure. *Annu. Rev. Biophys. Chem.* **18**, 47-69.
17. Klimov, D.K. & Thirumalai, D. (1996). Criterion that determines the foldability of proteins. *Phys. Rev. Lett.* **76**, 4070-4073.
18. Baumgartner, A. (1987). Simulations of polymer models. In *Applications of the Monte-Carlo Simulations in Statistical Physics.* pp. 281-319, Springer, New York.
19. Irbäck, A. & Schwarze, H. (1995). Sequence dependence of self-interacting random chains. *J. Phys. A: Math. Gen.* **28**, 2121-2132.
20. Berriz, G.F., Gutin, A.M & Shakhnovich, E.I. (1997). Cooperativity and stability in a Langevin model of protein like folding. *J. Chem. Phys.* **106**, 9276-9285.
21. Guo, Z., Brooks III, C.L. (1997). Thermodynamics of protein folding: a statistical mechanical study of a small all-β-protein. *Biopolymers* **42**, 745-757.
22. Zhou, Y., Karplus, M, Wichert, J.M & Hall, C.K. (1997). Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte-Carlo simulations of system with square-well interactions. *J. Chem. Phys.* **107**,10691-10708.
23. Zhou, Y. & Karplus, M (1997). Folding thermodynamics of a three-helix-bundle protein. *Proc. Natl Acad. Sci. USA* **94**, 14429-14432.
24. Doi, M (1996). *Introduction to Polymer Physics.* Clarendon Press,

Oxford.
25. Shakhnovich, E.I. & Gutin, A.M. (1993). Engineering of stable and fast folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
26. Abkevich, V.I. Gutin, A.M. & Shakhnovich, E.I. (1996). Improved design of stable and fast-folding model proteins. *Fold. Des.* **1**, 221-230.
27. Alder, B.J. & Wainwright, T.E. (1959). Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **31**, 459-466.
28. Grosberg, A. Yu. & Khokhlov, A.R. (1997). *Giant Molecules.* Appendix, Academic Press, Boston.
29. Allen, M.P. & Tildesley, D.J. (1987). Molecular dynamics. In *Computer Simulation of Liquids.* pp. 102-110, Clarendon Press, Oxford.
30. Rapaport, D.C. (1997). Step potential. In *The Art of Molecular Dynamics Simulation.* pp. 285-316, Cambridge University Press, Cambridge.
31. Mirny, L.A., Abkevich, V. & Shakhnovich, E.I. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Fold. Des.* **1**, 103-116.
32. Landau, L.D. & Lifshiftz, E.M. (1980). *Statistical Physics.* Pergamon, London.
33. Wetlaufer, D.B. (1973). Nucleation, rapid folding, and globular interchain regions in proteins. *Proc. Natl Acad. Sci. USA* **70**, 691-701.
34. Karplus, M. & Weaver, D.L. (1976). Protein-folding dynamics. *Nature* **260**, 404-406.
35. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1995). Domains in folding of model proteins. *Protein Sci.* **4**, 1167-1177.
36. Lazaridis, T. & Karplus, M. (1997). "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278**, 1928-1931.
37. Anifsen, C.B. (1973). Principles that govern the folding of the protein chains. *Science* **181**, 223-230.
38. Tsong, T.Y. & Baldwin, R.L. (1978). Effects of solvent viscosity and different guanidine salts on the kinetics of ribonuclease A chain folding. *Biopolymers* **17**, 1669-1678.
39. Bai, Y., Sosnick, T.R., Mayne, L. & Englander, S.W. (1995). Protein folding intermediates: native-state hydrogen exchange. *Science* **269**, 192-197.
40. Lacroix, E., Bruix, M, López-Hernández, E., Serrano, L. & Rico, M (1997). Amide hydrogen exchange and internal dynamics the chemotactic protein CheY from *Escherichia coli*. *J. Mol Biol.* **271**, 472-487.
41. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **34**, 827-828.
42. Brooks, C.L. III. (1992). Characterization of "native" apolyglobin by molecular dynamics simulations. *J. Mol. Biol.* **227**, 375-380.
43. Daggett, V. & Levitt, M (1992). A model of the molten globule state from molecular dynamics simulations. *Proc. Natl Acad. Sci. USA* **89**, 5142-5146.
44. Sheinerman, F.B. & Brooks, C.L. III. (1997). A molecular dynamics simulation study of segment B1 of protein G. *Proteins* **29**, 192-202.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad —for further information, see the explanation on the contents pages.