Multiple Folding Mechanisms of Protein Ubiquitin

Jian Zhang,[1] **Meng Qin,**[1] **and Wei Wang**[1,2*]
[1]*National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, China*
[2]*Interdisciplinary Center of Theoretical Studies, Chinese Academy of Sciences, Beijing, China*

ABSTRACT **Based on the C_α Go-type model, the folding kinetics and mechanisms of protein ubiquitin with mixed α/β topology are studied by molecular dynamics simulations. The relaxation kinetics shows that there are three phases, namely the major phase, the intermediate phase and the slowest minor phase. The existence of these three phases are relevant to the phenomenon found in experiments. According to our simulations, the folding at high temperatures around the folding transition temperature T_f is of a two-state process, and the folding nucleus is consisted of contacts between the front end of α-helix and the turn$_4$. The folding at low temperature (∼ $T = 0.8$) is also studied, where an A-state like structure is found lying on the major folding pathway. The appearance of this structure is related to the stability of the first part (residue 1–51) of protein ubiquitin. As the temperature decreases, the formation of secondary structures, tertiary structures and collapse of the protein are found to be decoupled gradually and the folding mechanism changes from the nucleation–condensation to the diffusion–collision. This feature indicates a unifying common folding mechanism for proteins. The intermediate phase is also studied and is found to represent a folding process via a long-lived intermediate state which is stabilized by strong interactions between the β_1 and the β_5 strand. These strong interactions are important for the function of protein ubiquitin as a molecular chaperone. Thus the intermediate phase is assumed as a byproduct of the requirement of protein function. In addition, the validity of the current Go-model is also investigated, and a lower limited temperature for protein ubiquitin $T_{limit} = 0.8$ is proposed. At temperatures higher than this value, the kinetic traps due to glass dynamics cannot be significantly populated and the intermediate states can be reliably identified although there is slight chevron rollover in the folding rates. At temperature lower than T_{limit}, however, the traps due to glass dynamics become dominant and may be mistaken for real intermediate states. This limitation of valid temperature range prevents us to reveal the burst phase intermediate in the major folding phase since it might only be stabilized at temperatures lower than T_{limit}, according to experiments. Our works show that caution must be taken when studying low-temperature intermediate states by using the C_α Go-models. Proteins 2005;59:565–579.**

Key words: validity of Go-model; intermediate states; glass dynamics; folding mechanisms

INTRODUCTION

The folding kinetics and folding mechanisms are central problems for the study of protein folding both experimentally and theoretically. Theoretical methods vary from minimalist models of protein, including lattice and off-lattice simulations,[1–13] to all-atom models.[14–22] Although simulations based on all-atom models provide much detailed information of folding, it can only resolve the time scale of several nanoseconds in one run or reach several microseconds combining a large number of runs. To completely characterize the nature of the energy landscape and the kinetics of folding, ensemble averaging over simulations or long time-running is still quite difficult and beyond the current computer capacity. Thus the minimalist models are the main tools for studying the energy landscape and folding mechanisms of proteins.[23]

Among the minimalist models, the Go-type models have shown some distinguished success in protein folding and have been widely used.[6–10] The studies on two-state folders (such as Cl2, SH3, barnase, and Im9[7,8,24] and three-state folders (such as CheY, Rnase, and Im7[8,9,24]) have shown that the folding kinetics and thermodynamics, as well as the overall structures of transition states and intermediates can be modelled successfully by the Go-type models, and the folding rates can also be obtained qualitatively.[10] The main reason of such success is due to that the real protein sequences are sufficiently well optimized or designed by nature and the folding mechanisms are mostly dominated by the native structures and the compensation between energy and entropy of the chains.[6–8,25–27]

Beyond the two-state or three-state folding mechanism, proteins may folds by multiple mechanisms simultaneously. For example, for a small three-helix-bundle protein with 46 residues, it was found that for large-gap models, the protein can fold simultaneously using a two-state mechanism and a three-state mechanism with two nonobligatory intermediates.[28,29] For small-gap models,

Grant sponsor: the National Natural Science Foundation of China; Grant numbers: 90103031, 90403120, 10021001, and 10474041; Grant sponsor: the Nonlinear Project (973) of the NSM.

*Correspondence to: Wei Wang, National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, 210093, China. E-mail: wang.wei@nju.edu.cn

Received 18 May 2004; Accepted 1 November 2004

Published online 14 March 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20430

however, the mechanism becomes three-state and one of intermediates becomes obligatory. Other examples include a designed four-helix bundle,[5,30,31] lysozyme,[32] DHFR,[7] cytochrome c_{551},[33] the FEP WW domain,[34] and a proline-free variant of staphylococcal nuclease,[35] and so on. All these proteins show complex kinetics with multiple folding mechanisms, a fraction of molecules reaches the native state via a two-state mechanism, whereas the rest pass through one or several intermediates. This has been termed as kinetic partition mechanism and was proposed as a unifying scheme in the folding of biomolecules by Thirumalai et al.[5,30,31] The reason comes from the polymeric nature and the presence of conflicting energy scales in proteins, and also that the free energy landscape is rough and contains not only the native basin attraction, but competing basins of attraction as well. The rough free energy landscape results in direct and indirect pathways to the native basin, that is, a kinetic partitioning mechanism.

The work by Karplus and coworkers on a small protein with a three-helix-bundle showed that for this small protein with rather simple topology multiple folding mechanisms and the structures of intermediates can be described quite well by the Go-type model related to the experimental findings. Now the question is: can the Go-type models characterize the folding behaviors of larger proteins with complex topology and multiple folding mechanisms, can the partition among different mechanisms and the related intermediate states be predicted correctly? Previously, a Go-type model has been used to study the folding of protein DHFR[9] which is a two-domain 162-residue α/β enzyme. It folds via intermediate I_{HF} represented as a set of structures I_1–I_4 which are structurally similar to each other but proceed toward the native state with different rates. The study based on the Go-model for this protein showed that a set of intermediates can be seen and the overall structures of the intermediates are in agreement with the experiment findings. However, the partition among different folding pathways and the corresponding folding rates have not been studied. Although proteins with multiple folding mechanisms have long been discovered both theoretically[5,28–30,36–38] and experimentally,[31–35] detailed Go-type modelling studies on multiple folding mechanisms of topological complex proteins are rare.

As a small globular protein with 76 residues, ubiquitin shows unusual native structural features and folding kinetics. As shown in Figure 1, protein ubiquitin includes a five-strand β-sheet with three antiparallel and one parallel pairs of strands, an α-helix (residues 23–34), two short helices (helix-1, residues 38–40; helix-2, residues 56–59), and seven reverse turns. Among the turns, five locate near the contact region between the front end of α-helix and turn-4 (see Fig. 1).[39] The curved β-sheet and the flanking α-helix enclose a single core of densely packed hydrophobic side chains that contributes to the high structural stability of ubiquitin. The protein has important functions and acts as a chaperone for proteasomal degradation.[40]

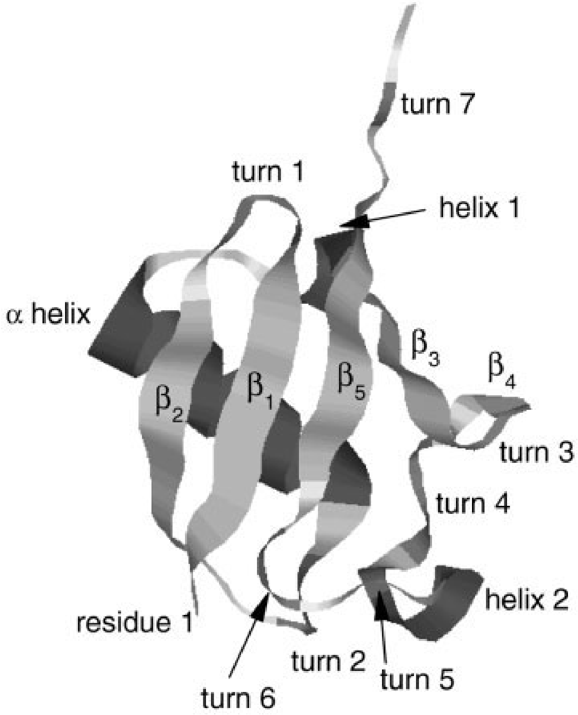

Fig. 1. The 3D structure of the protein-ubiquitin (PDB code: 1ubq). The figure is ploted by RasMol.

The hydrogen exchange and stopped-flow fluorescence experiments on ubiquitin refolding have shown a relaxation kinetics with three distinct time scales:[41–44] a fastest major folding phase, an intermediate phase and a slowest minor phase. Experimentally, the major phase has aroused many debates where the central question is whether there is an intermediate state during the major folding process.[41–44] In recent years, due to the advances of experimental techniques, this question is becoming clear. There is indeed an hidden intermediate during folding, but it can only be observed under strong native conditions, such as at low temperatures or with presence of stabilizing additives.[45–47] Besides, for the protein ubiquitin, the nature of the intermediate phase is not very clear either. As for the slowest minor phase, it is attributed to the *cis–trans* isomerization of proline residues. Therefore experiments have suggested that protein ubiquitin reaches its native state via different pathways with different mechanisms. The hidden intermediate state in the major phase and the multiple folding pathways of ubiquitin provide a real challenge to the Go-models. Can the Go-models successfully describe the folding of this protein?

Theoretically, there has been many works on the folding of ubiquitin. These works include study of backbone desolvation and mutational hot spots of ubiquitin,[48] study of the folding nucleus, structure of the transition state, folding mechanisms and pathway heterogeneity using a Ramachandran basin folding algorithm,[49,50] all-atom molecular dynamical simulation of the hydrophobic collapse and the A-state of protein ubiquitin,[51,52] the study of the conservation of folding nucleus residues in ubiquitin superfamily[53] and so on. Especially, a designed ubiquitin-like protein with 68 residues is studied based on the so-called BLN model.[38] In that work, multiple exponential kinetics was observed. The burst phase intermediate was also observed, but it has little secondary structures, which agrees with previous experiments[54] whereas it contradicts with a recent CD experiment.[45] Although there has been much progress, the studies aimed to reveal hidden intermediate states in the folding of ubiquitin are still lacking, and the possibly different folding mechanisms at different temperatures are still not fully understood.

To help understand the folding of the protein ubiquitin and to check how far the C_α Go-type models can go for such proteins with multiple folding pathways, a Go-type model is employed to study the folding of protein ubiquitin. First, the valid temperature range of the current Go-model is studied, and a lower limit T_{limit} is found. Within the valid temperature range, different relaxation kinetics and folding mechanisms at high and low temperatures are investigated. The structures of intermediate state responsible for the intermediate phase and the stabilizing interactions of this intermediate state are also studied. At last, the slowest minor phase found in our simulations is studied. Based on these studies, a discussion for the advantages and limitations of the C_α Go-models is made.

THE MODEL

The folding of the protein ubiquitin is studied by using an off-lattice C_α based the Go-type model. In the model, all residues are represented as beads centered in their C_α positions, and interact with each other by bond, angle, dihedral angle, and 10–12 Lennard-Jones interactions. The Hamiltonian is shown in Eq. (1).[8]

$$U = \sum_{\text{bonds}} K_r(r - r_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedral}} K_\phi^{(n)} \times \{1 - \cos[n(\phi - \phi_0)]\} + \sum_{i<j-3} \left\{ \epsilon(i,j)\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right] + \epsilon_2(i,j)\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}\right\}. \quad (1)$$

In Eq. (1), r and r_0 represent the distances between two subsequent residues at the calculated conformation and native state, respectively. Analogously, $\theta(\theta_0)$ and $\phi(\phi_0)$ represent the corresponding angle and dihedral angle, respectively. The last term contains the Go-type nonbonded interactions, $\epsilon(i,j) = \epsilon$ and $\epsilon_2(i, j) = 0$ if residues i and j form a native contact, while $\epsilon(i, j) = 0$ and $\epsilon_2(i, j) = \epsilon$ if they do not. Residues i and j are assumed to form a "native contact" if the distance between any two heavy-atoms that belong to these two residues is within 5 Å in the native state. In the simulations, residues i and j are assumed to form a "contact" if their distance is within 1.2 times of their native distance. The parameter σ (i, j) is taken equal to the distance between two C_α atoms in residues i and j at the native state for native contacts, while $\sigma(i, j) = 4$ Å for the other pairs. Parameters are taken to be $K_r = 100\epsilon$, $K_\theta = 20\epsilon$, $K_\phi^{(1)} = \epsilon$ and $K_\phi^{(3)} = 0.5\epsilon$, respectively.

The simulation package AMBER (version 7.0) is used to do the simulations[55] at constant temperature (NTB = 0, NTT = 1, see AMBER's user guide). The Berendsen algorithm is invoked to couple the system to an external bath, and the coupling constant is chosen as *tautp* = 1.0 (its unit is the same as that of time, see below). The time unit (t.u.) is arbitrary defined as 500 MD steps, that is if we say current time is $t = 1$ *t.u.,* that means 500 MD steps. This is just for expression simplicity and is compatible with AMBER.

In the simulation, the native structure of the protein is first heated to temperature $T = 2.0$ (the folding temperature is $T_f = 1.07$ in our model, which is obtained from the main peak in temperature dependence of specific heat C_v curve) to unfold them to denatured states with $Q < 0.2$. Here Q is the fraction of the total native contacts. Then the temperature is jumped to the study temperature and the protein begins to refold. The molecules are assumed to be folded when the reaction coordinate $Q \geq 0.95$ and the radius of gyration $R_g \leq 12$ Å (for comparison, the R_g of native state calculated based on the C_α atoms is 11.8 Å). The refolding processes are simulated at several temperatures ranging from $T = 0.7$ to $T_f = 1.07$. At each temperature, more than 1000 independent trajectories are collected except at temperature $T = 0.8$, where more than 8000 trajectories are calculated.

RESULTS AND DISCUSSION

Relaxation Kinetics

The time evolution of population of denatured states $P(t) = \int_t^\infty \cdot f(t')dt'$ is monitored to study the refolding kinetics. Here $f(t)$ is the distribution of first passage time (FPT) for reaching the native state with $Q \geq 0.95$ and $R_g \leq$ 12.0 Å. The ensemble averaged $\langle P(t)\rangle$ at temperatures from $T = 0.7$ to $T = 1.0$ are shown in Figure 2. For the temperatures close to the folding temperature T_f, such as $T = 1.0$, the value of $\langle P(t)\rangle$ decreases exponentially. In contrast, at low temperature, such as $T = 0.8$, $\langle P(t)\rangle$ first decreases exponentially with a fast relaxation time constant and 90% molecules fold within time $t = 100$*t.u.,* which is called the major folding phase hereafter (see inset of Fig. 2). After the major folding phase, $\langle P(t)\rangle$ decreases exponentially by a slower time constant, called the intermediate phase. Besides these two phases, another minor folding phase can also be seen at $T = 0.8$ with a slowest time constant (see the tail of the curve). Thus the folding relaxation processes can be divided into three phases, a major phase with fastest relaxation constant, an intermediate phase, and a slowest minor phase.

It is found that $\langle P(t)\rangle$ can be fitted quite well with the following formula,

$$\langle P(t)\rangle = A_0 + A_1\exp(-(t/\tau_1)^\beta) + A_2\exp(-t/\tau_2) + A_3\exp(-t/\tau_3). \quad (2)$$

The fitting based on this equation leads to reduced error $\chi^2 \sim 10^{-6}$ at all temperatures. The obtained parameters are given in Table I. Note that a similar formula with only two exponential terms will result in a larger error, whereas

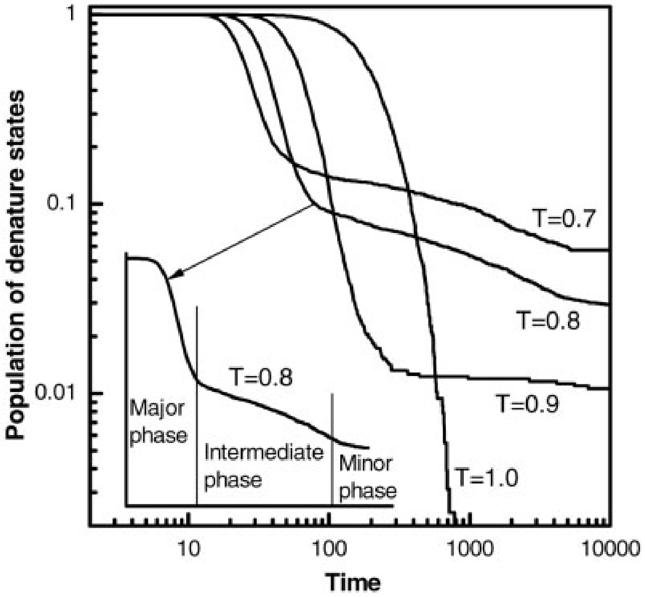

Fig. 2. Refolding relaxation monitored by time evolution of populations of denatured states at temperatures from $T = 0.7$ to $T = 1.0$. Both the x and y coordinates are logarithmic ones. At high temperature ($T = 1.0$), the relaxation kinetics shows a single exponential phase. However, at low temperature it shows three exponential phases, a fastest major phase, an intermediate phase and a slowest minor phase, which can be seen clearly in the inset.

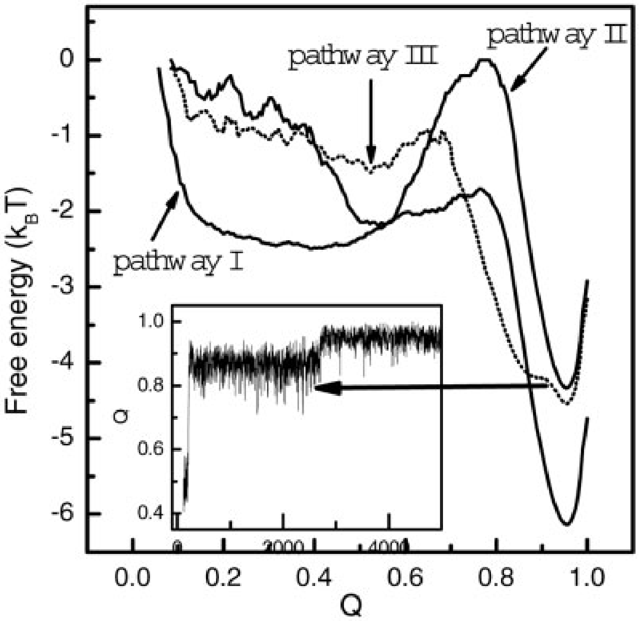

Fig. 3. The free energy as a function of Q of the three folding pathways which are related to three phases respectively. The inset shows a typical trajectory which folds through the pathway III, it illustrates that the intermediate and native basin overlap each other and the molecule has to use a long time to overcome the folding barrier.

TABLE I. Fitting Parameters for the Kinetic Traces in Figure 2

T	A_0	A_1	τ_1	β	A_2	τ_2	A_3	τ_3
0.7	0.06	0.78	27.2	4.1	0.12	49.6	0.08	1405.9
0.8	0.03	0.86	39.1	3.87	0.09	80.8	0.04	1683.9
0.9	0.01	0.92	70.2	3.05	0.09	106.6	0.00	5400.3
1.0	0	0.95	233.9	2.02	0.08	327.9	0	—
1.07	—	1	1377	1.34	0	—	0	—

two relaxation time constants τ_i will be same if four exponential terms are employed. As proved that the kinetics can be divided into three phases. The A_1, A_2, and A_3 terms are related to the major phase, the intermediate phase and the slowest minor phase, respectively. Except for refolding through these three phases, about 2% of trajectories do not reach their folded state within the upper limit of simulation time $t_{max} = 10000t.u.$. However, with sufficiently large t_{max}, we believe that they will fold to the native state finally. It should be noted that at all temperatures below the folding temperature $T_f = 1.07$, the parameter β obtained from our simulations is larger than 1. It is due to the strong bias of free energy surface toward the native state at these temperatures, and the smooth feature of the energy funnel of the Go-type model. With increasing temperature, the free-energy surface is biased against native state gradually, β decreases and reaches 1 at the temperature slightly above the folding temperature. This behavior has been reported recently in a study of two-dimensional lattice HP model.[56]

The temperature dependence of the amplitudes A_i and the folding time constants τ_i of three phases are shown in Table I. At low temperatures, all the amplitudes A_1, A_2, and A_3 are non-zero. As the temperature increases, the amplitude A_1 of the major phase increases monotonically and the amplitude A_2 and A_3 decrease monotonically. When the temperature approaches T_f, the major phase becomes dominant and the relaxation curve shows a single exponential form. As for the folding time constants, both the time constants of the major (τ_1) and the intermediate phases (τ_2) increase as the temperature increases, but the increasing speed of the intermediate phase is smaller than that of the major phase. Then, these two time constants coincide at the temperature T_f, and two processes merge there. Note that the time constants of the slowest minor phase λ_3 are always very large at all temperatures.

The fact that the population of denatured states relaxes with a three-exponential kinetics reflects that the protein folds through three different pathways with different folding mechanisms. Hereafter, we label these three pathways as I, II, and III, corresponding to the major phase, the intermediate phase, and the slowest minor phase, respectively. This feature has long been discovered by Thirumalai et al. and named kinetic partitioning mechanism.[31] The free energy of these three pathways calculated at $T = 0.8$ is given in Figure 3. To calculate this figure, more than 8000 folding trajectories are collected and the conformations are classified by their Q values, then the free energy is calculated as negative logarithm of the population, that is, $F = -\log\langle P(Q)\rangle$, where $P(Q)$ is the population of the states with their fraction of native contacts are Q. Such kinds of free energy are not the free energy in equilibrium states, but they can also illustrate the relative population of states during the folding process; in this sense, they are also free energy. When calculating Figure 3, the trajectories are partitioned into three classes, each corresponds to a folding phase. The trajectories with their FPT smaller than $t = 39t.u.$ are assumed passing through pathway I and corresponding to the major phase. The trajectories that reach the native basin within the time range $81t.u. < t < 300t.u.$ are assumed to be related to the intermediate phase. The choice of limitation of $t = 300t.u.$ is somewhat arbitrary, but the calculated quantities are not sensitive to this choice since very few trajectories in the intermediate phase have an FPT larger than this value. Analogously, the trajectories with their FPT larger than $t = 1700t.u.$ are assumed belonging to the slowest minor phase. Note that all the free energy plots in this article are calculated in a similar way as mentioned above, thus the calculating method will be not repeated later.

According to Figure 3, the free energy barrier for folding along pathway I is about $1k_BT$, this small barrier is consistent with the fast-folding rate of the major phase. The rate limiting barrier for folding along pathway II (located at $Q = 0.75$) is about $2.5k_BT$, resulting in a slower folding rate of the intermediate phase. The free energy well located at $Q = 0.5$ in the pathway II corresponds to an intermediate state, which will be discussed later. For the pathway III, the free energy curve is somewhat confusing at first sight since there is no obvious large barrier responsible for the slowest folding rate of the minor phase. In fact, the rate limiting barrier located at $Q = 0.9$ is a very large barrier. It is not obvious at first sight because the intermediate basin prior to this barrier and the native basin overlap each other when projected on the reaction coordinate Q. This barrier is actually very high, which can be seen clearly from the inset of Figure 3. The inset shows a typical trajectory, which folds through the pathway III. It clearly shows that the intermediate and native basin overlap each other and the molecules have to use a long time to jump out the intermediate basin and enter the native basin. It can be expect that this barrier will be obvious when the free energy is projected on an appropriate reaction coordinate. This implies that one should be very cautious when choosing reaction coordinates since incorrect reaction coordinates may lead to wrong conclusions.

The Valid Temperature Range of the Current Go-model and Intermediate States

Although the Go-type models have been very successful in understanding the folding of many small proteins, it was pointed out by Chan and Kaya recently that the common Go-models fall short of producing the type of calorimetric two-state cooperativity observed for many small proteins.[57–61] For the common Go-models, the chevron rollovers emerge under strongly native conditions due to a competition between a stronger driving force for folding and the onset of glass dynamics under strongly native conditions (large values of ϵ/kT). As the native condition becomes stronger (for examples, when temperature is lowered), the kinetic trapping becomes more prominent, the protein has to use more time to overcome the barrier to proceed folding. This raises a question on the validity of the common Go-models at low temperatures, especially when one tries to find the hidden intermediates that can only be populated at low temperatures. Under low temperatures, the kinetic traps due to the glass dynamics may be mistaken for real intermediates. Thus it is very important to investigate the valid temperature range of the model before studying the folding mechanism.

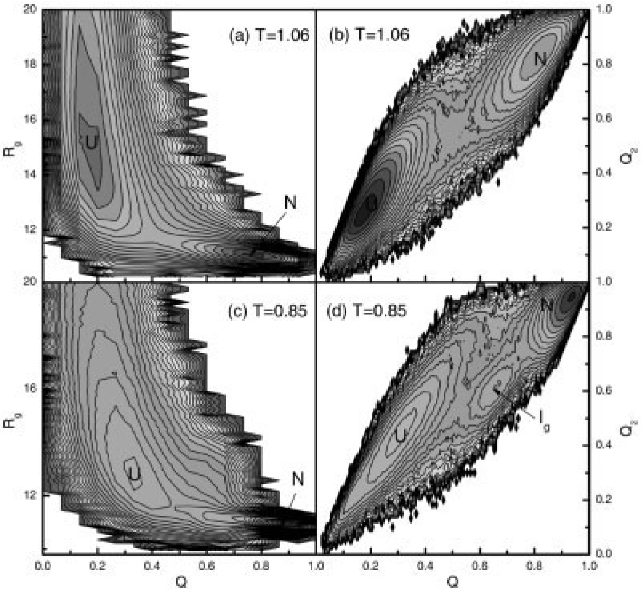

Fig. 4. The free energy contour plots for protein CI2 projected onto reaction coordinates Q-R_g (**a**) and Q-Q_2 (**b**) at temperature $T = 1.06$, slightly above the folding temperature $T_f = 1.05$. The Q are the fraction of total native contacts, Q_2 the fraction of native contacts in secondary structures, and R_g the radius of gyration, respectively. (**c**) and (**d**) are the same plots but calculated at temperature $T = 0.85$. The free energy difference between adjacent contour lines is $0.5k_BT$. The unfolded basin, native basin and an intermediate basin are marked by U, N, and I_g, respectively.

To do this, we have calculated the free energy of ubiquitin as a function of temperature. For comparison, the results of protein CI2 (PDB code: 1coa) are given first. Figure 4 shows the free energy contour plots of protein CI2 projected onto reaction coordinates Q-R_g [Fig. 4(a)] and Q-Q_2 [Fig. 4(b)] at the temperature $T = 1.06$ and $T = 0.85$, respectively. Here Q is the fraction of native contacts, R_g is the radius of gyration and Q_2 is the fraction of contacts in secondary structures. The free energy is calculated as negative logarithm of the population, for example, $F = -log[P(Q, R_g)]$, where $P(Q, R_g)$ is the population of the states with their fraction of native contacts and radius of gyration are Q and R_g, respectively. Each plot is obtained by averaging on more than 1000 trajectories.

As shown in Figure 4(a, b), at the temperature $T = 1.06$, slightly above the folding temperature ($T_f = 1.05$), there are only two basins populated: the native (N) and unfolded basin (U). At low temperature $T = 0.85$, however, a third basin I_g appears in the Q-Q_2 plot, as shown by Figure 4(d). From experiments, it is known that there is no intermediate state for protein CI2. Thus the appearance of this basin I_g can only be attributed to the glass dynamics of the

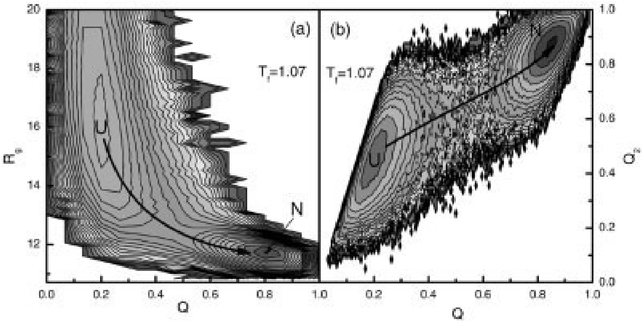

Fig. 5. The free energy contour plots for protein ubiquitin projected onto reaction coordinates Q-R_g (**a**) and Q-Q_2 (**b**) at the folding temperature T_f = 1.07. The free energy difference between adjacent contour lines is $0.5k_BT$.

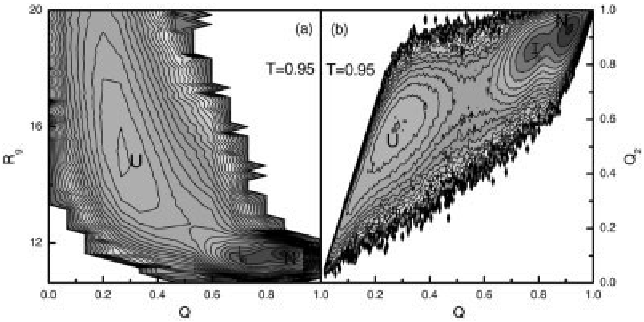

Fig. 6. The free energy contour plots for protein ubiquitin projected onto reaction coordinates Q-R_g (**a**) and Q-Q_2 (**b**) at temperature T = 0.95. An intermediate basin I_m begins to appear at this temperature. The free-energy difference between adjacent contour lines is $0.5k_BT$.

model. This indicates that the present Go-model has a rather high glass temperature $T_g \sim 0.85$. Actually, it is this kind of basin that contributes to the chevron rollover of the folding rates, which was first related to the glass dynamics by Chan and Kaya.[57,60] Therefore, for CI2, the valid temperature of the C_α Go-model should be limited to $T > 0.85$. Any observation below this temperature should be carefully checked to ensure that it is not caused by the glass dynamics. It is also worth noting that the basin I_g cannot be observed in Q-R_g plot, the free energy projections onto these two reaction coordinates are not as sensitive as the Q-Q_2 plot.

The folding of protein ubiquitin is much complex compared with protein CI2. Figures 5–8 show the free energy contour plots for protein ubiquitin at four temperatures T_f = 1.07, 0.95, 0.8, and 0.75, respectively. When calculating these figures, all trajectories of three folding phases are used. As expected, the protein shows two-state behavior at the folding temperature T_f (Fig. 5), only unfolded basin U and native basin N are populated. This is consistent with experiments under weak native conditions.[44,47] Another populated basin I_m is first observed at $T = 0.95$ (Fig. 6), as can be seen in both Q-R_g and Q-Q_2 contour plots of the free energy. This intermediate basin cannot be assumed as traps due to the glass dynamics since its first appearance is at the temperature $T = 0.95$. If we estimate the folding temperature for ubiquitin to be roughly 350 K, the temperature $T = 0.95$ corresponds to 311 K. Thus, the states that emerge at such a high temperature can hardly be attributed to the glass dynamics. In fact, this basin is related to the slowest minor phase in relaxation kinetics, as will be discussed later.

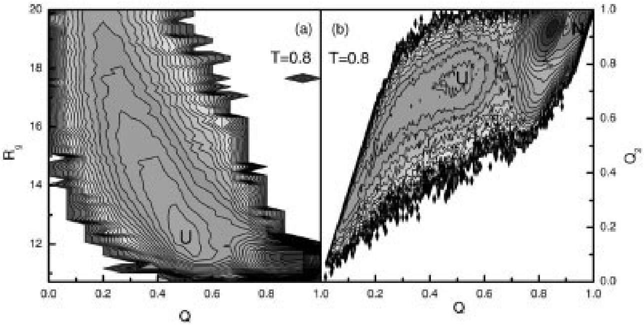

Fig. 7. The free energy contour plots for protein ubiquitin projected onto reaction coordinates Q-R_g (**a**) and Q-Q_2 (**b**) at temperature T = 0.8. Note the U basin shifts greatly to the native side, which signals the existence of a hidden intermediate state. The free-energy difference between adjacent contour lines is $0.5k_BT$.

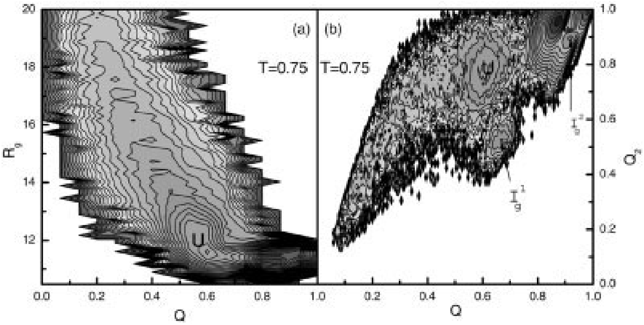

Fig. 8. The free-energy contour plots for protein ubiquitin projected onto reaction coordinates Q-R_g (**a**) and Q-Q_2 (**b**) at temperature T = 0.75. Two traps, I_g^1 and I_g^2 begin to appear at this temperature which are related to the glass dynamics. The Q-R_g plot (**a**) is insensitive compared to the Q-Q_2 plot (**b**). The free energy difference between adjacent contour lines is $0.5k_BT$.

When the temperature is lowered to $T = 0.75$, as shown in Figure 8(b), at least two new basins emerge, namely I_g^1 and I_g^2. These two basins only appear at temperature lower than $T = 0.75$ (corresponds to roughly 245 K), thus it is highly possible that they are caused by the glass dynamics of the model.

For temperatures $T \geq 0.8$, as shown in Figures 5–7, except the above mentioned basin I_m, there are only two main basins populated, i.e., U and N basins. From these figures, it is very interesting to note that the U basin shifts greatly to the native side as the temperature decreases. At temperature $T = 0.8$, more than 70% secondary structures have been formed for the states in this basin [see the Q_2 coordinate of the U basin in Figure 7(b)] and the radius of gyration of the geometry center of the basin is 12.3 Å [Fig. 7(a)], close to the value of the native state. This raises such questions as: Is the U basin at low temperature only a shifted unfolded basin or a signature of a hidden intermediate state? If the later case is true, does this intermediate state arise from the glass dynamics?

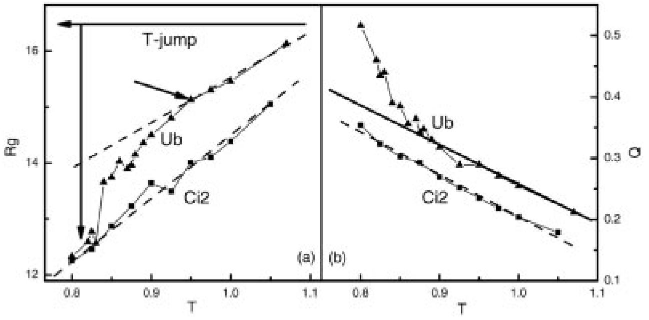

Fig. 9. The shift of the geometry center of the U basin as a function of temperature for protein CI2 and ubiquitin, illustrated by the value of R_g (**a**) and Q (**b**) respectively. The straight lines for CI2 is the best fit, the lines for ubiquitin are only a guide for the eye.

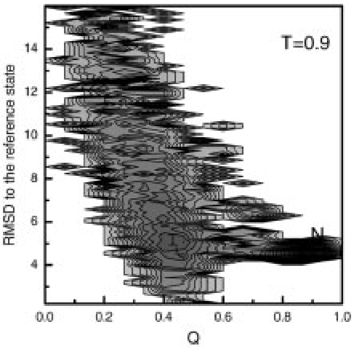

Fig. 10. The free energy contour plot for the intermediate phase projected onto Q-RMSD at temperature $T = 0.9$. The RMSD is calculated by using a representative intermediate state as reference state. An intermediate basin is revealed and marked by I_i. The free-energy difference between adjacent contour lines is $0.5k_BT$.

To answer such questions, the shift of the geometry centers of the U basin of both protein CI2 and ubiquitin as a function of temperature are calculated and shown in Figure 9. The difference between two proteins can be seen clearly. In case of CI2, the values of R_g (Q) decrease (increase) linearly as temperature decreases, this demonstrates that the U basin at low temperature for protein CI2 is merely a shifted unfolded basin. For protein ubiquitin, however, the behavior of R_g (Q) begins to deviate from linearity at temperature $T = 0.95$ and the values of R_g (Q) decrease (increase) rapidly as the temperature decreases further. When such a protein is subject to a T-jump experiment, according to Figure 9(a), the R_g related signal will show a burst phase and the amplitude of the burst phase as a function of temperature will be a sigmoid-shape. Experimentally, such a behavior of the signal is often interpreted as the existence of a hidden intermediate state. Thus it is highly possible that the great shift of the U basin at lower temperature for ubiquitin indicates the existence of a hidden intermediate state.

By carefully checking the trajectories of three phases, we found that there is indeed a hidden intermediate state I_i which is responsible for the intermediate phase in relaxation kinetics. The great shift of the U basin for protein ubiquitin in Figure 9 is only superficial and due to partial overlap of the unfolding basin of the major phase and the intermediate basin I_i when projecting the free energy onto the reaction coordinates. To prove such an argument, we select a representative state I_{ref} within the I_i basin and calculate the Q-RMSD contour plot of the free energy for the intermediate phase. The RMSD is calculated by taken I_{ref} as a reference state. When calculating these plots, only the trajectories passing through the pathway II are used. Here, which pathway the trajectories will pass through is determined by their FPT (see the folding time constants in Table I and the previous discussions). The result at $T = 0.9$ is shown in Figure 10 where the free energy difference between adjacent contour lines are $0.5k_BT$. An intermediate basin I_i can be seen clearly. If the I_i basin is defined by the edge whose free energy is $1k_BT$ higher than that of the center, the width of the basin is W (RMSD) ≤ 2 Å and $W(Q) \sim 0.1$. Such relatively small variance of RMSD and Q suggest that the states within this basin are highly structured, thus the I_i basin is indeed an intermediate basin instead of an unfolded basin. This proves our argument that the great shift of the unfolded basin U at low temperature implies the existence of an hidden intermediate state. This intermediate cannot be attributed to the glass dynamics of the model since Figure 10 shows that it can be populated at a high temperature such as $T = 0.9$ and its effect can be firstly seen at $T = 0.95$ [Fig. 9(a)]. Here $T = 0.95$ corresponds to 294 K if T_f is estimated to be about 350 K for protein ubiquitin. On the other hand, the intermediate I_i cannot be caused by the Go-type model itself since there is no evidence of intermediate state at such high temperatures for protein CI2 with exactly the same model (Fig. 9).

Thus, by comparing with the results for protein CI2, we conclude that the valid temperature range for the present model should be limited to $T \geq T_{limit} = 0.8$ since at $T = 0.75$ several traps due to the glass dynamics have been observed. For temperatures higher than $T = 0.8$, although there is slight chevron rollover in the folding rates, which indicates the onset of the glass dynamics, the kinetic traps cannot be significantly populated at these relatively high temperatures and the real intermediate states can still be identified. Differently, for temperatures below the T_{limit}, the glass dynamics become dominant, and the real intermediate states will be difficult to distinguish from the traps caused by the glass dynamics.

The Major Phase

Experimentally, the major phase [the A_1 term in Eq. (1)] is the fastest one among three phases and accounts for large percent of total amplitude change of signals during folding process. The nature of the major phase has arisen many debates, such as whether the folding of the major phase is a two-state or three-state process, whether an intermediate state exists and has some structures, and so on. Recently, due to the advances in experimental techniques, it has become clear that there is indeed a hidden intermediate state for ubiquitin, but this intermediate state may only exist under strong native conditions, such as at low temperatures or with stabilizing additives.[45–47]

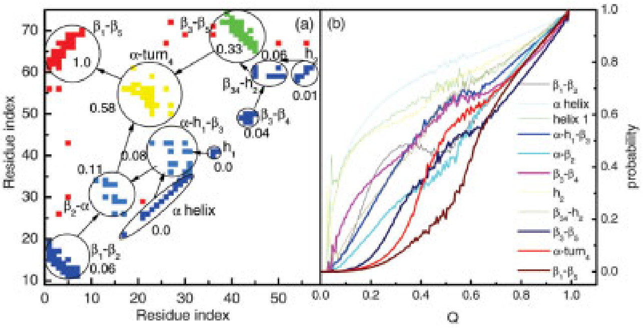

Fig. 11. (**a**) The formation order of structural elements for protein ubiquitin calculated at $T = 1.07$. The contacts are colored by their first formation time, the contacts with blue color form earlier, and the ones with red color form later. For clarity, the first formation time of each contact cluster has been rescaled to the range [0,1] and labelled beside it. The arrows indicate the folding routes. (**b**) The formation probability of each contact cluster as a function of *Q*. The important contact clusters are plotted by thicker lines.

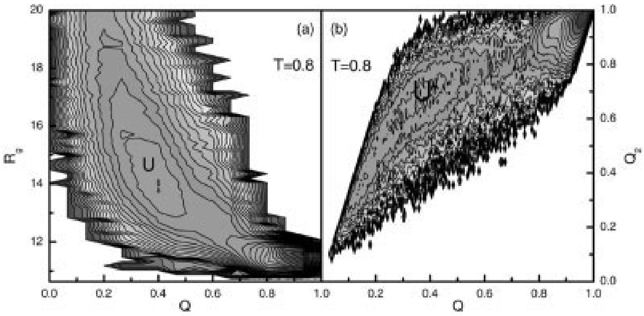

Fig. 12. The free-energy contour plot for the major phase calculated at temperature $T = 0.8$. The free-energy difference between adjacent contour lines is $0.5k_BT$. It should be noted that the *U* basin is rather broad and shallow.

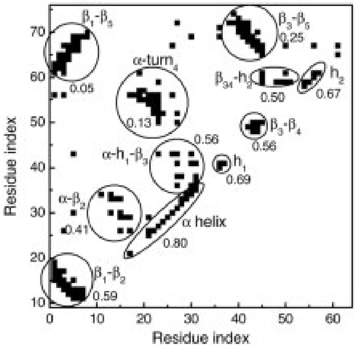

Fig. 13. The average structure of the states in the *U* basin in Figure 12, showing that these states are partially structured. The formation probability of each contact cluster is given by the number beside it.

In the following section, we investigate the folding behaviors at high and low temperatures and hope to understand the folding of protein ubiquitin. At the same time, we check the prediction abilities of the current Go-model concerning the intermediate states.

Folding of ubiquitin at high temperature

The contour plots of the free energy of the major phase at the folding temperature $T_f = 1.07$ $(\leq T_f)$ have been shown

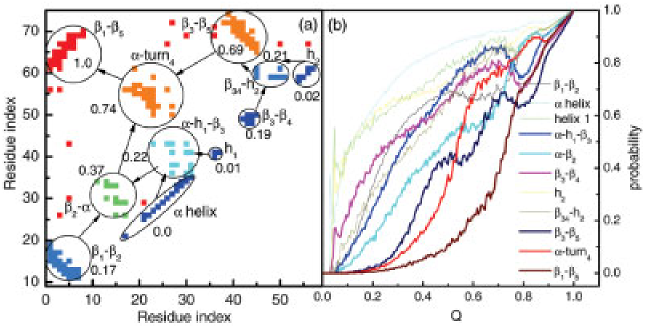

Fig. 14. (**a**) The formation order of structural elements for protein ubiquitin calculated at $T = 0.8$. The contacts are colored by their first formation time. The first formation time of each contact cluster has been rescaled to the range [0,1] and labelled beside it. The arrows indicate the folding routes. (**b**) The formation probability of each contact cluster as a function of *Q*. The important contact clusters are plotted by thicker lines.

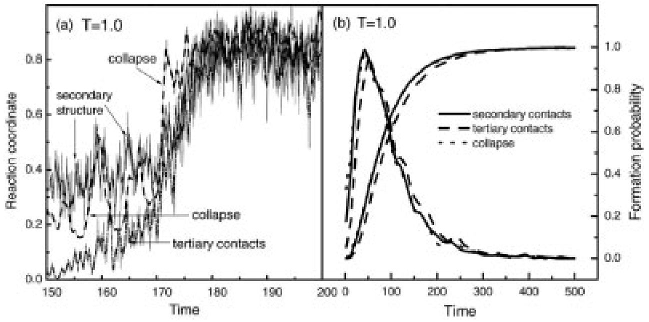

Fig. 15. (**a**) Time evolution of the secondary structure, tertiary contacts, and the "collapse" of a typical trajectory at temperature $T_f = 1.0$. The "collapse" is calculated by $R_g^3(N)/R_g^3$, where the $R_g(N)$ is the radius of gyration of the native state. (**b**) The average formation probability of the secondary structure and tertiary contacts and the collapse of protein as a function of time, the three curves with peaks at the left side are the corresponding derivatives. These curves are obtained by averaging more than 1000 trajectories.

in Figure 5. Two populated basins can be seen in the figure, labelled by *U* and *N*, respectively. At the folding temperature, protein ubiquitin is clearly a standard two-state folder. Figure 5(b) also shows that the secondary and tertiary structures form cooperatively since the most possible folding pathway marked by the black line lies approximately on the diagonal. These results are in agreement with the observations in hydrogen exchange experiments,[42,54] stopped-flow fluorescence intensity[44] and far-UV CD spectroscopy measurements.[54] In these experiments, the burst phase intermediate has not been detected, suggesting that the secondary structural elements can be populated only marginally ahead of the major cooperative folding events. This supports the two-state folding mechanism with cooperative formation of the secondary and tertiary structures.

The detailed folding orders of the structural elements for ubiquitin at $T = 1.07$ is illustrated in Figure 11. Figure 11(a) shows the first formation time of each structural element and Figure 11(b) gives their formation probabilities as a function of *Q*. Note that both of them were averaged on more than 1000 trajectories. To calculate

Figure 11(b), all the conformations during the folding process are classified by their Q values, and the formation probability of each contact at Q is calculated by the number of conformations with this contact formed divided by the total number of conformations with this Q. The formation probability of each contact cluster is obtained by averaging on all contacts in the cluster. Figure 11(a) and 11(b) are combined to describe the overall folding process of protein ubiquitin. In Figure 11(a), the contacts are colored by their first formation time, the contacts with blue color form earlier, and the ones with red color form later. The first formation time of each contact cluster is also labelled beside it (the time has been rescaled to the range [0,1], see the time labelled beside each contact cluster). It can be seen that the secondary structures form first, such as the α-helix, helix-1, helix-2, the $\beta_1\beta_2$ hairpin, and the $\beta_3\beta_4$ hairpin. Then the tertiary contacts between the central α-helix, helix-1 and β_3 are formed. At almost the same time, the contacts between $\beta_3\beta_4$ hairpin and helix-2 are formed. Shortly after, the contacts between α-helix and β_2 are also formed. It should be noted that the formation of these contacts does not necessarily mean that they will not break later. Actually these contacts keep fluctuating, i.e., forming and breaking, which can be seen from the fluctuation of formation probabilities in Figure 11(b). We also emphasize that in Figure 11 we intend to show the relative formation order of structural elements. In fact, the formation of different structural elements is rather cooperative for the folding at temperatures close to T_f, as will be discussed in detail later.

From Figure 11(a), the formation of contacts between α-helix and turn-4 occurs rather late compared to the fast formation of secondary structures and some tertiary contacts. Thus it can be reasonably assumed that the formation of these contacts is the rate-limiting step for protein folding. That is, these contacts serve as candidates of critical contacts for protein folding, namely, the folding nucleus. By watching movies of the folding process, it is also found that only until these contacts are formed, are the two parts of protein pulled together and the molecules overcome the transition state and reach their native state.

The folding nucleus can be calculated by checking several long-time trajectories at the folding temperature T_f and picking out the FF and UU conformations containing ample information of transition state. Here the FF conformations are those ones that originate in and return to folded region without ascending to the unfolded region and the UU conformations originate in and return to the unfolded region without descending to the folded region.[62] The nucleus is constructed by these contacts that appear much more often in the FF conformations than in the UU conformations (for the detailed description of this method, refer to a work done by Shakhnovich's group[62]). For protein ubiquitin, we found by using this method that the folding nucleus is composed of the contacts between the front part of α-helix and turn-4. This is consistent with our analysis based on the formation orders of structural elements in Figure 11(a).

The position of the folding nucleus can explain why a *cis-trans* isomerization of PRO-19 affects the folding of the entire molecule whereas the isomerization of PRO-37 or PRO-38 does not propagate into other parts of the molecule.[43] The reason is that the PRO-19 locates very close to the nucleus in both sequence and space. The presence of a *cis* peptide bond close to PRO-19 will inhibit the formation of folding nucleus at this region, therefore the molecule cannot overcome the transition state and has to return back to unfolded basin.

Figure 11(a, b) also shows that the last folding event is the packing of the β_5 strand onto the hydrophobic core which is composed of the central α-helix and the curved β-sheet. After the packing, the contacts between β_3 and β_5, β_1 and β_5 are formed. Then the entire β sheet is constructed. This folding picture agrees with the previous one suggested based on experiments[43] or simulation results.[38] There the first formation of the tertiary structure occurs cooperatively between the central α-helix and β-sheet, and the formation of the C-terminus loop region is the last event in the folding process.

Folding of ubiquitin at low temperature

Recent experiments on ubiquitin have shown that the burst phase intermediate is only marginally stable at room temperature and in water solution. However, it can be stabilized under strong native conditions, such as at low temperature[45,46] or with high concentration of stabilizing additives.[47] This prompts us to calculate the free energy of the major folding phase at low temperature since it yields a strong native condition.

According to previous discussions, the valid temperature range of current Go-model is $T \geq T_{limit} = 0.8$. For temperatures lower than T_{limit}, the glass dynamics becomes dominant and the resulting traps may be mistaken for intermediate states. Therefore, the low-temperature behavior of ubiquitin folding is studied at temperature $T = 0.8$. The free energy plots of the major phase projected on Q-R_g and Q-Q_2 are calculated at this temperature and shown in Figure 12(a, b). To calculate these figures, only the trajectories with their FPT smaller than the relaxation constant of the major phase ($\tau_1 = 39t.u.$) are used. The reason of such selection is that the molecules folding along different pathways sample different part of free energy landscape. According to Figure 12, there are only two basins populated, labelled by U and N respectively. The U basin also shifts to the native side comparing to its position at high temperatures, similar to the behavior of the U basin in Figure 7. However, for the major phase, the R_g of the geometry center of the U basin is linearly dependent on the temperature, indicating that the shift of the U basin does not signal an intermediate state. Further, if we define the U basin as the region circled by the contour line with its free energy $1k_BT$ higher than that of the center, the width of this basin is W(R_g) ~ 4 Å and W(Q) ~ 0.3 (this value corresponds to 57 contacts), as shown by Figure 12(a). The large width of the basin suggests that the fluctuation of states in the U basin is large and the states are only partially structured.

The average structure of the states located in the U basin at temperature $T = 0.8$ is shown in Figure 13. It will be easier to understand the structure if we divide the protein into two parts: the first one comprises the $\beta_1\beta_2$ hairpin, the central helix, helix$_1$, and the $\beta_3\beta_4$ hairpin (residues 1–51) and the second part is composed of the turn-4, helix-2, and the β_5 strand (residues 51–76). This division is based on the contact map in Figure 13 which shows that the contacts inside the first part are all well formed (such as the $\beta_1\beta_2$ hairpin, $\beta_3\beta_4$ hairpin, the central helix and helix$_1$, contacts between β_2 and α-helix, and between α-helix, helix$_1$, and β_3 strand), whereas the contacts between these two parts are basically not formed (such as the contacts between β_3 and β_5, between α-helix and turn-4, and between β_1 and β_5). The structure of the states in the U basin implies that the the first part of ubiquitin is rather stable, and can exist weakly dependent on the second part. These states are similar to the A-state in protein ubiquitin which is formed at PH 2.0 in 60% methanol/40% water.[63] The A-state is characterized by the presence of first β-hairpin and part of the third strand, the hydrophobic face of the β-sheet is covered by a partially structured α-helix. The structure of the A-state is consistent with our argument that the first part of protein ubiquitin is relatively stable. Such a stability is also relevant to the experimental results on peptide fragments of ubiquitin (each fragment with residue 1–51).[64] Two fragments form a dimer (the S state) that is stabilized by 0.8 M sodium sulfate at room temperature. The structure of each fragment and the interface between two fragments mimic related features in the structure of intact ubiquitin.

Although the states in the U basin are partially structured, we prefer to assume this U basin as an shifted unfolded basin instead of an intermediate state. This is due to the fact that the width of this basin is rather broad (see Fig. 12) and the R_g of the geometry center of the U basin is linearly dependent on the temperature. These partially structured states are not necessarily contradictory to the concept of unfolded states since the unfolded states are found with some native-like mean structures.[65] It is interesting to note that the width of this basin will continuously decrease at much lower temperature. For example, at $T = 0.6$, the width of the basin becomes $R_g \sim 2$ Å and $Q \sim 0.2$. Unfortunately, this temperature goes beyond the valid temperature range of the model and it is hard to distinguish the hidden intermediate states from traps due to the glass dynamics. Experimentally, the burst-phase intermediate state can only be observed at low temperatures ($T < 4°\text{C}$).[46] If we estimate the folding temperature of protein ubiquitin to be 350 K, this corresponds to $277/350 \sim 0.79$, right below the temperature limitation $T = 0.8$ of the current Go-model. This low temperature prevents us from revealing the burst phase intermediate observed in the major phase of ubiquitin folding. Although we cannot reproduce the hidden intermediate observed in experiments at low temperatures due to the limitation of current model, we found that the states in the U basin become more and more structured as the temperature decreases. Their structures resemble the structures of the A-state of ubiquitin, which was suggested to be on the folding pathway of ubiquitin.[64] Whether the A-state-like structures are related to the burst phase intermediate state is a very interesting question and needs further study.

The folding mechanism at low temperatures is different from that at high temperatures. The detailed folding order of the structural elements at temperature $T = 0.8$ is shown in Figure 14. Figure 14(a) shows the first formation time of each structural element and Figure 14(b) gives their formation probabilities as a function of Q during the folding process. Figure 14(a) is very similar to Figure 11(a) but the formation times of structural elements are different. At high temperatures, the formation of contacts are more “compact” in time, thus the folding has higher cooperatively comparing to that at low temperature. This can be seen more clearly by comparing Figure 14(b) with Figure 11(b). In the former case most curves lie at the diagonal and close to each other. In the later case, however, the curves depart from each other, indicating that the formation of different structural elements decouples somewhat. The difference between Figure 14 and Figure 11 indicates that there exist different folding mechanisms at different temperatures. Actually, according to our simulations, the folding mechanism at high temperatures is a nucleation–condensation one, it will slide to a diffusion–collision mechanism as the temperature decreases, this feature will be discussed in detail in the following section.

The folding mechanisms at high and low temperatures

At high temperature ($T \sim T_f$), the folding mechanism is a nucleation–condensation one which involves cooperative formation of the secondary and the tertiary structures. To demonstrate such a mechanism more clearly, Figure 15(a) shows a typical trajectory calculated at $T = 1.0$. Before the time $t = 170t.u.$, the trajectory is characterized by a long-time search for the correct nucleus, indicated by the fluctuation of the inverse fraction of native volume (see the curve marked with “collapse”). Little stable secondary and tertiary structures are formed during this period. After $t = 170t.u.$, the correct nuclei are formed, then the molecules undergo a suddenly collapse and the secondary and tertiary structures are formed cooperatively, manifesting a nucleation– condensation mechanism. The average formation probabilities of the secondary and tertiary contacts and the collapse of protein as a function of time are shown in Figure 15(b), which is obtained by averaging on more than 1000 trajectories. The three curves with peaks at the left side are their respective derivatives, showing the distribution of formation time of secondary and tertiary structures and collapse of protein. It is clearly that the three events occur closely to each other in time, manifesting a high folding cooperativity and the nucleation–condensation mechanism.

At low temperatures ($T < T_f$), however, the folding mechanism becomes diffusion–collision-like gradually. Figure 16(a) gives the same plot as Figure 15(a) but calculated at $T = 0.8$. At $t = 17t.u.$, almost all the secondary

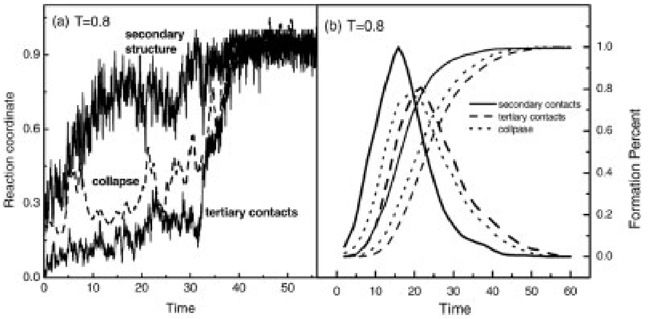

Fig. 16. (**a**) Time evolution of the secondary structure, tertiary contacts, and collapse of the protein of a typical trajectory at temperature T = 0.8, similar to Figure 15. (**b**) The average formation probability of the secondary structure, tertiary contacts and collapse of protein and their respective derivatives as a function of time, same as Figure 15, but calculated at temperature $T = 0.8$.

structures (> 80%) are formed, whereas less than 30% tertiary contacts are formed and the volume of the molecule is about three times larger than the native volume. These are the typical aspects of the diffusion–collision mechanism. After $t = 17t.u.,$ the molecule keeps searching for the correct packing of formed structural elements in a slow diffusion–collision-like manner until mission accomplished at $t = 32t.u.$. Then the molecule collapses to the correct conformation and the formation of tertiary contacts follows. The average formation probability of the secondary structure, the tertiary contacts and the collapse of protein are shown in Figure 16(b). In contrast to the case at high temperatures [Fig. 15(b)], the collapse of protein and formation of the tertiary contacts clearly lag behind the formation of the secondary structure, suggesting a low folding cooperativity and diffusion–collision mechanism.

Previously, such sliding from one mechanism to another under different conditions has been seen in protein three-helix studied by a Go-model.[28] It has been found that under strong bias gap (optimized protein), the protein folding mechanism tends to be a diffusion–collision one, whereas for less optimized protein, the mechanism changes to one involving simultaneous collapse and partial secondary structure formation, followed by reorganization to the native structure. This picture is similar to the behavior of ubiquitin in our simulations. In our case, the temperature acts as a parameter like bias gap in their study. This is due to the fact that generally lower temperature leads to a stronger bias of free energy surface to the native state whereas high temperature results in a weaker bias toward the native state or even against it.

In our opinion, the sliding from nucleation–condensation to diffusion–collision mechanism with increasing of native conditions (by decreasing temperature or using some other techniques) may be a general feature in protein folding. Physically, the formation of the secondary structures should occur prior to the major folding event due to their small size. But for many proteins, the stability of the secondary structures are low and cannot be populated without the protection by the tertiary structures, thus the MG state with ample secondary structures cannot be observed in experiments. This low stability of secondary structures demands the proteins to fold cooperatively and leads to a nucleation–condensation mechanism. However, if the stability of secondary structures increases, such as by simulating at low temperatures, the formation of the secondary and tertiary structures may be decoupled and the folding mechanisms will slide to the diffusion–collision one. In this sense, the two folding mechanisms are different facets of an unifying mechanism. The key to which facet a protein manifests is the relative stability of the secondary and tertiary structures, which can be tuned by temperature, solvent additives, mutation of the sequence, or other techniques. This unifying mechanism has also been reported by recent work that shows that the mechanisms can slide from nucleation–condensation to diffusion–collision/framework for proteins in a superfamily of three-helical.[66]

We also emphasis that the C_α-based Go-model can only describe the general features of protein folding mechanisms instead of predicting the behaviors of the specific protein such as ubiquitin. In our opinion, similar remarks should also be made for the previous work on three-helix.[28] This is because for real proteins, the stability of a secondary structure may be mainly determined by the hydrogen bonds, the strong native interactions between side chains,[67] non-native docking of structural elements[68] or some other interactions, whereas the effect of the temperature in the experimental range may only serve as a small perturbation. Thus we can reasonably imagine that only one folding mechanism can be observed for these proteins. In the current model, the sequence information and side chain effects as well as the non-native interactions are omitted, so it cannot be used to predict the folding mechanism of specific protein such as ubiquitin or three-helix, the results should be viewed as a general feature of protein folding.

The Intermediate Phase and the Slowest Minor Phase

The folding of the protein ubiquitin is rather complex since it folds not only through the major phase but also through another slow intermediate phase. This intermediate phase has long been observed experimentally and it accounts for up to 20% of the amplitude.[44] The intermediate phase is generally attributed to the *cis-trans* isomerization of the proline residues. However, according to the experiments,[44] its rate constant is within a factor of five of the rate constant of the major phase. Since the rate constant of the major phase measured in stopped-flow experiments at 0.54 M GdmCL solvent condition is 368s^{-1},[44] the rate constant of the intermediate state is estimated to be 70s^{-1}. We think that this rate appears to be too fast for proline isomerization.[41,43] Furthermore, the intermediate phase still exists even in double-jump experiments where the isomerization effects have been weakened. Thus the origin of the intermediate phase is very interesting and will be discussed in this section.

In our simulations, the intermediate phase is observed in the relaxation kinetics and is found to be related to the intermediate state I_i, as has been discussed above. In Figure 10 we have given the free energy of the intermedi-

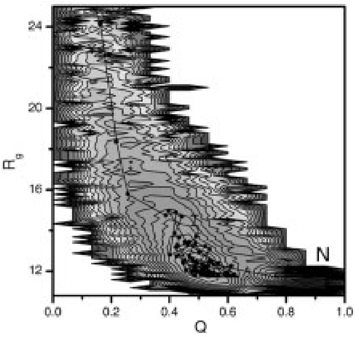

Fig. 17. The free energy contour plot projected on reaction coordinates Q-R_g for the intermediate phase. Also shown is a typical trajectory which clearly shows that the intermediate state I_i is an on-pathway one.

ate phase projected on Q-RMSD reaction coordinates to show the existence of the intermediate state I_i. In Figure 17 we give the free energy contour plot projected on Q and R_g at temperature $T = 0.8$, also shown is a typical trajectory which folds through this phase (see the connected lines). To calculate this free energy plot, only trajectories which reach the native state between time $t = 81t.u.$ and $t = 300t.u.$ are used since the relaxation constants of the intermediate phase and the slowest minor phase are $\tau_2 = 81t.u.$ and $\tau_3 = 1684t.u.$ respectively. The choice of $t = 300t.u.$ is somewhat arbitrary but does not affect the results significantly, since very little trajectories have an FPT larger than this value. The intermediate basin I_i represents an on-pathway intermediate since the trajectories need not return back to the unfolded states before reaching the native basin. According to Figure 17, the barrier separating the I_i state from the native basin is about $2.5k_BT$. The center of the intermediate basin I_i locates at $R_g \sim 12.5$ and $Q \sim 0.5$. From the value of R_g, this intermediate state is very compact, almost as compact as the native state [$R_g\,(native) = 11.8$ Å]. These suggests that the intermediate I_i is rather stable. This intermediate state cannot be attributed to the glass dynamics or the artifact of the Go-model, as have been discussed above.

By monitoring the conformational motions of the intermediate state I_i, several transient structures can be found and these structures inter-convert quickly between each other (the movie not shown here). A common feature of these transient structures is that all of them seem to be stabilized by the interactions between β_1 and β_5 strands. Only after the contacts between two strands are broken, can the molecules escape from the intermediate basin I_i and proceed folding. Further evidence for this argument can be seen in Figure 18(a) which shows the averaged contact map over states in the intermediate basin I_i. A remarkable feature of the contact map is that the formation probability of contacts between β_1 and β_5 is even higher than that between the central helix and turn-4. Recall that the contacts between β_1–β_5 should be formed at the final stage of folding whereas the contacts between α-helix and turn-4 are formed at the transition state. The formation of contacts for the latter should be earlier, than that for the former. Thus the contacts between β_1 and β_5 shown in Figure 18(a) are actually "misfolded contacts" which have to be broken later for the molecules to escape from the intermediate. This can be seen quite clearly in Figure 18(b) which gives the formation probabilities of all tertiary contacts as a function of Q for the intermediate folding phase. In that figure only the formation probabilities of the contacts between β_1 and β_5 strands drop before the transition state (locating at $Q = 0.75$, see Fig. 3), showing that these contacts must be broken first to proceed folding.

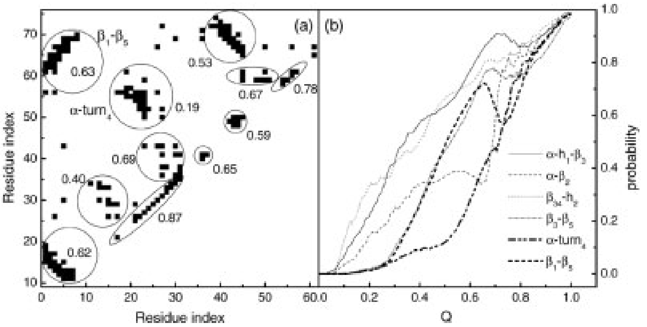

Fig. 18. (**a**) The average contact map of states locating in the intermediate basin I_i in Figure 17. The formation probability of each contact cluster is given by the number labelled beside it. Note that the formation probability of contacts between β_1 and β_5 is even larger than that between α-helix and turn-4. (**b**) The formation probabilities of all tertiary contacts as a function of Q. The secondary structures are not plotted since they cannot be the stabilizing factors of the intermediate I_i. The figure shows that only the formation probabilities of contacts between β_1 and β_5 strands drop before the transition state.

The stabilizing role of the interactions between β_1 and β_5 strands can also be seen from simulations based on mutations of the related residues. In the mutations, after an arbitrary contact between β_1 and β_5 is eliminated, the amplitude of the intermediate phase is decreased, indicating the intermediate state is destabilized. It is also found that if the interaction strengths of contacts between β_1 and β_5 are decreased by a factor of 2, the intermediate phase will totally disappear. Therefore, the stabilizing factor of the intermediate state is attributed to the strong interactions between β_1 and β_5 strands. Such strong interactions in this region are also supported by an experimental measurement of the force between these two strands using a single-molecule force spectroscopy technique.[40]

Taking consideration of the fact that these strong interactions are required by the biological function of the protein ubiquitin as a chaperone for proteasomal degradation,[40] the intermediate phase observed in experiments is assumed to be a byproduct of the requirement of the protein's biological function. Previously, Fersht pointed out that:[69] "Although there is evolutionary pressure on proteins to fold rapidly, folding rates do have to fit in with the functional requirements of each particular protein. Because of this, it is possible that many proteins may have to fold by slower mechanisms because they carry portions of essential structure that slow down folding. Thus, inherently slower mechanisms in which intermediates accumulate will also be observed." This explains the origin of the intermediate phase observed in the protein ubiquitin's folding. Here the functional requirements on the protein structures yield a slower folding pathway for protein ubiquitin, and such a folding pathway manifests itself as a slower intermediate phase in experiments.

Beside the major and the intermediate phases, we also observed a slowest minor phase in the relaxation kinetics. This phase takes less than 4% of all trajectories and its relaxation time constant is much larger than that of the major and intermediate phase. As discussed above, this phase is linked to the intermediate basin I_m in Figures 6 and 7. By carefully checking the structures of states in this basin, we find that all of them are with their N-terminus or C-terminus buried in the hydrophobic core (figure not shown). This cannot be possible for real proteins, since the side chains will overlap each other and cause extremely high potential energy for these conformations, thus the I_m states are not realistic in our opinion. The appearance of the I_m states should be attributed to the lack of side-chain effects of the current C_α model. It can be expected that this minor phase and the related intermediate state I_m will disappear in an all-atom Go-model.[70]

Conclusions

Our work is aimed to undertake a systematic understanding of the folding of the protein ubiquitin with complex topology, and to see how-far the Go-type model can go for such kinds of proteins with multiple folding pathways.

By comparing the results with those of protein CI2, we found that the application of the current model should be limited to a temperature range $T \geq T_{limit}$. For ubiquitin, T_{limit} is roughly 0.8. At temperatures higher than this limitation, although there are slightly chevron rollovers in the folding rates, the traps due to the glass dynamics cannot be significantly populated and the real intermediate states can be reliably identified if they exist. At temperatures lower than the limitation, the glass dynamics become dominant. Although we may still observe real intermediates, it is hard to distinguish them from traps due to the glass dynamics. Because of this limitation of the current C_α Go-model, the burst intermediate in the major folding phase of the protein ubiquitin cannot be reproduced in our simulations since this intermediate may only exist at very low temperatures. However, we find that the folding mechanism at low temperatures is different from that in high temperatures, and an A–state-like state lies on the folding pathway at low temperatures. Whether this state is related to the burst-phase intermediate is interesting and needs to be further studied.

This limitation of Go-type model is due to lack of many body interactions in the model according to Chan et al. [60,61] As Chan has pointed out,[61] to ascertain the physical bases of the many-body interactions, it would be extremely interesting to see how side-chain packing, hydrogen bonds, and other atomic interactions may give rise to the mechanisms of local–nonlocal coupling. In our opinion, one important factor omitted in current C_α Go-model is the effect of side chains. At temperatures lower than T_{limit}, the traps due to the glass dynamics dominate the folding and cause serious chevron rollovers in the arms of folding rates. For real proteins, these trap states may be destabilized by the interactions between the side chains which will cause high-energy contacts thus eliminating these traps, so only the native and real intermediate states have reasonable energy and can be populated. (This can be partly seen by the observation of an intermediate I_m which we think will disappear when the effect of side chains is introduced into the model). This destabilization of kinetic traps is consistent with the requirement of high cooperativity that "entails not only the stabilization of the native structure but also the destabilization of otherwise stable nonnative conformations".[61] By this way the interactions between side chains may serve as one of the physical bases of the high folding cooperativity and the extremely low glass temperature T_g of real proteins. (A recent work suggests that for real calorimetrically two-state proteins, the energy landscape theory "folding to glass transition temperature ratio" T_f/T_g may exceed 6.0.[58]) Presumably, during the folding process, the native interactions guide the protein to native state while the side-chain interactions pave the road by stuffing the possible kinetic traps. When the side-chain effect is introduced into the current model, the glass temperature may be greatly decreased, then the model can be used at much lower temperatures and the hidden intermediate in ubiquitin folding observed in experiments at low temperature may be reproduced in simulations. Of course, our arguments here are highly tentative and need to be further investigated.

It is very interesting to note that in a very recent work, the c-Crk SH3 domain has been studied over a broad range of temperatures by using a C_β Go-model. Below the kinetic partition temperature T_{KP}, two intermediates have been found. However, since the main conclusions in this work are deduced from simulations at $T = 0.33$, its ratio over T_f is about 0.53, the detected intermediate states in this protein have to be examined carefully by experiments.

Although the burst phase intermediate state of the major phase cannot be reproduced in the current model, we do find an intermediate state I_i which cannot be attributed to the glass dynamics or the artifact of the model itself. The reason is that this intermediate state can be populated at a high temperature and at this temperature no intermediate states can be observed in protein CI2 by using exactly the same model. This is consistent with the success of the Go-type models reported previously on the reproduction of intermediates and transition states. The success of the Go-models suggests that native interactions are the primary determinant of most protein folding, and that non-native interactions lead only to local structural perturbations. Although recent literature shows that some on-pathway intermediates are actually misfolded species, there are also reports manifesting that some intermediates are mainly stabilized by the native interactions and the non-native interactions are not essentially for the formation of these intermediates.[71] Remarkably, the Go-models also demonstrate consistency with the general features of transition states though to be stabilized by non-native interactions.[72] These manifest that the

Go-type models are still powerful even when non-native interactions are presented but not dominant.

In summary, both the limitation and success of the current C_α Go-model have been presented in this work. Generally speaking, the C_α Go-type models can still be good tools to study protein folding if we are only interested in the transition states, intermediate states or some other properties under weakly native conditions, such as at not too-low temperatures relative to T_f. In this area, the success of the Go-models is surprising. The model's success comes from that most proteins are highly optimized and their folding are mainly determined by the native interactions and that non-native interactions lead only to local structural perturbations. After introducing side chains of residues into C_α model, the all-atom Go-model should be workable in much larger areas. It can be expected that, even for more complex problems, such as interactions between proteins and interactions between proteins and other biomolecules, e.g., DNA and biological membranes, the Go-type model and its more realistic versions will still be a powerful method. These will enable us to have a deeper understanding of the mystery of the structure and function of biomolecules.

ACKNOWLEDGMENTS

We thank two anonymous referees and the editorial manager for their critical reading of the manuscript and very helpful comments.

REFERENCES

1. Gō N. Theoretical studies of protein folding. Annu Review Biophys Bioeng 1983;12:183–210.
2. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
3. Sali A, Shakhnovich EI, Karplus M. How does a protein fold. Nature 1994;369:248–251.
4. Shakhnovich EI, Farztdinov G, Gutin AM, Karplus M. Protein folding bottlenecks: a lattice monte carlo simulation. Phys Rev Lett 1991;67:1665–1668.
5. Guo Z, Thirumalai D. Kinetics and thermodynamics of folding of a de novo designed four-helix bundle protein. J Mol Biol 1996;263: 323–343.
6. Nymeyer H, Carcía AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc Natl Acad Sci USA 1998;95:5921–5928.
7. Clementi C, Jennings PA, Onuchic JN. Prediction of folding mechanism for circular-permuted proteins. J Mol Biol 2001;311: 879–890.
8. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298: 937–953.
9. Clementi C, Jennings PA, Onuchic JN. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1β. Proc Natl Acad Sci USA 2000;97:5871–5876.
10. Koga N, Takada S. Roles of native topology and chain-length scaling in protein folding: a simulation study with a Gō-like model. J Mol Biol 2001;313:171–180.
11. Wang J, Wang W. Folding transition of model protein chains characterized by partition function zeros. J Chem Phys 2003;118: 2952–2963.
12. Qin M, Wang J, Tang Y, Wang W. Folding behaviors of lattice model proteins with three kinds of contact potentials. Phys Rev E 2003;67:061905(1–8).
13. Li J, Wang J, Zhang J, Wang W. Thermodynamic stability and kinetic foldability of a lattice protein model. J Chem Phys 2004;120: 6274–6287.
14. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
15. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci USA 1998;95: 9897–9902.
16. Daggett V, Levitt M. A model of the molten globule state from molecular dynamics simulations. Proc Natl Acad Sci USA 1992;89: 5142–5146.
17. Daggett V. Long timescale simulations. Curr Opin Struct Biol 2000;10:160–164.
18. Petrella RJ, Karplus M. A limiting-case study of protein structure prediction: energy-based searches of reduced conformational space. J Phys Chem 2000;104:11370–11378.
19. Snow CD, Nguyen N, Pande VS, Gruebele M. Absolute comparison of simulated and experimental protein-folding dynamics. Nature 2002;420:102–106.
20. Kussell E, Shimada J, Shaklmovich EI. A structure-based method for derivation of all-atom potentials for protein folding. Proc Natl Acad Sci USA 2002;99:5343–5348.
21. Shen MY, Freed KF. All-atom fast protein folding simulations: the villin headpiece. Proteins 2002;49:439–445.
22. García AE, Onuchic JN. Folding a protein in a computer: an atomic description of the folding/unfolding of protein A Proc Natl Acad Sci USA 2003;100:13898–13903.
23. Head-Gordon T, Brown S. Minimalist models for protein folding and design. Curr Opin Struct Biol 2003;13:160–167.
24. Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE. Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. J Mol Biol 1999;286: 1597–1608.
25. Takada S. Gō-ing for the prediction of protein folding mechanisms. Proc Natl Acad Sci USA 1999;96:11698–11700.
26. Baker, D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
27. Muńoz V, Eaton WA. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;96:11311–11316.
28. Zhou YQ, Karplus M. Interpreting the folding kinetics of helical proteins. Nature 1999;401:400–403.
29. Zhou YQ, Karplus M. Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. J Mol Biol 1999; 293:917–951.
30. Veitshans T, Klimov D, Thirumalai D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. Fold & De 1996;2:1–22.
31. Thirumalai D, Klimov DK, Woodson SA. Kinetic partitioning mechanism as a unifying theme in the folding of biomolecules. Theor Chem Acc 1997;96:14–22.
32. Wildegger G, Kiefhaber T. Three-state model for lysozyme folding: triangular folding mechanism with an energetically trapped intermediate. J Mol Biol 1997;270:294–304.
33. Gianni S, Travaglini-Allocatelli C, Cutruzzolâ F, Brunori M, Ramachandra Shastry MC, Roder H. Parallel pathways in cytochrome $c_5$51 folding. J Mol Biol 2003;330:1145–1152.
34. Karanicolas J, Brooks III, CL. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: lessons for protein design? Proc Natl Acad Sci USA 2003;100:3954–3959.
35. Kamagata K, Sawano Y, Tanokura M, Kuwajhna K. Multiple parallel-pathway folding of proline-free staphylococcal nuclease. J Mol Biol 2003;332:1143–1153.
36. Abkevich VI, Gutin AM, Shakhnovich EI. Free energy landscape for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. J Chem Phys 1994;101:6052–6062.
37. Dinner AR, Karplus M. The Thermodynamics and kinetı̈cs of protein folding: a lattice model analysis of multiple pathways with intermediates. J Phys Chem 1999;103:7976–7994.
38. Sorenson JM, Head-Gordon T. Toward minimalist models of large proteins: a ubiquitin-like protein. Proteins 2002;46:368–379.
39. Laub PB, Khorasanizadeh S, Roder H. Localized solution structure refinement of an F45W variant of ubiquitin using stochastic boundary molecular dynamics and NMR distance restraints. Protein Sci 1995;4:973–982.
40. Carrion-Vazquez M, Li HB, Lu H, Marszalek PE, Oberhauser AF,

Fernandez JM. The mechanical stability of ubiquitin is linkage dependent. Nat Struct Biol 2003;10:738–743.

41. Khorasanizadeh S, Peters ID, Butt TR, Roder H. Folding and stability of a tryptophan-containing mutant of ubiquitin. Biochemistry 1993;32:7054–7063.
42. Khorasanizadeh S, Peters ID, Roder H. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. Nat Struct Biol 1996;3:193–205.
43. Briggs MS, Roder H. Early hydrogen-bounding events in the folding reaction of ubiquitin. Proc Natl Acad Sci USA 1992;89: 2017–2021.
44. Krantz BA, Sosnick TR. Distinguishing between two-state and three-state models for ubiquitin folding. Biochemistry 2000;39: 11696–11701.
45. Qin Z, Ervin J, Larios E, Gruebele M, Kihara H. Formation of a compact structured ensemble without fluorescence signature early during ubiquitin folding. J Phys Chem B 2002;106:13040–13046.
46. Larios E, Li JS, Schulten K, Kihara H, Gruebele M. Multiple probes reveal a native-like intermediate during low-temperature refolding of ubiquitin J Mol Biol 2004;340:115–125.
47. Went HM, Benitez-Cardoza CG, Jackson SE. Is an intermediate state populated on the folding pathway of ubiquitin? FEBS Lett 2004;567:333–338.
48. Fernández A. Time-resolved backbone desolvation and mutational hot spots in folding proteins. Proteins 2002;47:447–457.
49. Fernández A, Colubri A. Pathway heterogeneity in protein folding. Proteins 2002;48:293–310.
50. Sosnick TR, Berry RS, Fernández A, Colubri A. Distinguishing foldable proteins from nonfolders: when and how do they differ? Proteins 2002;49:15–23.
51. Alonso DOV, Daggett V. Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60% methanol and in water. J Mol Biol 1995;247:501–20.
52. Alonso DOV, Daggett V. Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci 1998;7:860–874.
53. Michnick SW, Shakhnovich E. A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. Fold Des 1998;3:239–251.
54. Gladwin ST, Evans PA. Structure of very early protein folding intermediates: new insights through a variant of hydrogen exchange labelling. Fold Des 1996;1:407–417.
55. Case DA, Pearlman DA, Caldwell JW, Cheatham III TE, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, and others. AMBER 7, 2002, University of California, San Francisco.
56. Nakamura HK, Sasai M, Takano M. Squeezed exponential tial kinetics to describe a nonglassy downhill folding as observed in a lattice protein model. Proteins 2004;55:99–106.
57. Chan HS. Modeling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. Proteins 2000;40:543–571.
58. Kaya H, Chan HS. Polymer principles of protein calorimetric two-state cooperativity. Proteins 2000;40:637–661.
59. Kaya H, Chan HS. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? J Mol Biol 2003;326:911–931.
60. Kaya H, Chan HS. Origins of chevron rollovers in non-two-state protein folding kinetics. Phys Rev Lett 2003;90:258104.
61. Chan HS, Shimizu S, Kaya H. Cooperativity principles in protein folding. Methods Enzymol 2004;380:350–379.
62. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Identifying the protein folding nucleus using molecular dynamics. J Mol Biol 2000;296:1183–1188.
63. Harding MM, Williams DH, Woolfson DN. Characterization of a partially denatured state of a protein by two-dimensional NMR: reduction of the hydrophobic interactions in ubiquitin. Biochemistry 1991;30:3120–3128.
64. Bolton D, Evans PA, Stott K, Brondhurst RW. Structure and properties of a dimeric N-terminal fragment of human ubiquitin. J Mol Biol 2001;314:773–787.
65. Zagrovic B, Snow CD, Khaliq S, Shirts MR, Pande VS. Native-like mean structure in the unfolded ensemble of small proteins. J Mol Biol 2002;323:153–164.
66. Gianni S, Guydosh NR, Khan F, Caldas TD, Mayor U, White GWN, DeMarco ML, Daggett V, and Fersht AR. Unifying features in protein-folding mechanisms. Proc Natl Acad Sci USA 2003;100: 13286–13291.
67. Cochran AG, Skelton NJ, Starovasnik MA. Tryptophan zippers: stable, monomeric b-hairpins. Proc Natl Acad Sci USA 2001;98: 5578–5583.
68. Gorski SA, Duff CSL, Capaldi AP, Kalverdal AP, Beddard GS, Moore GR, Radford SE. Equilibrium hydrogen exchange reveals extensive hydrogen bonded secondary structure in the on-pathway intermediate of Im7. J Mol Biol 2004;337:183–193.
69. Fersht AR. Nucleation mechanisms in protein folding. Curr Opin Struct Biol 1997;7:3–9.
70. Clementi C, García AE, Onuchic JN. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. J Mol Biol 2003;326:933–954.
71. Mizuguchi M, Kobashigawa Y, Kumaki Y, Demura M, Kawano K, Nittal K. Effects of a helix substitution on the folding mechanism of bovine α-Lactalbumin. Proteins 2002;49:95– 103.
72. Karanicolas J, Brooks III CL. Improved Co-like models demonstrate the robustness of protein folding mechanisms toward nonnative interactions. J Mol Biol 2003;334:309–325.