PROTEINS: Structure, Function, and Bioinformatics 67:795–810 (2007)

Folding of Tandem-Linked Domains

E. Prabhu Raman,[1] Valeri Barsegov,[2] and Dmitri K. Klimov[1]*
[1]*Department of Bioinformatics and Computational Biology, George Mason University, Manassas, Virginia 20110*
[2]*Department of Chemistry, University of Massachusetts Lowell, Lowell, Massachusetts 01854*

ABSTRACT **One of the factors, which influences protein folding *in vivo*, is a linkage of protein domains into multidomain tandems. However, relatively little is known about the impact of domain connectivity on protein folding mechanisms. In this article, we use coarse grained models of proteins to explore folding of tandem-linked domains (TLD). We found TLD folding to follow two scenarios. In the first, the tandem connectivity produces relatively minor impact on folding and the mechanisms of folding of tandem-linked and single domains remain similar. The second scenario involves qualitative changes in folding mechanism because of tandem linkage. As a result, protein domains, which fold via two-state mechanism as single isolated domains, may form new stable intermediates when inserted into tandems. The new intermediates are created by topological constraints imposed by the linkers between domains. In both cases tandem linkage slows down folding. We propose that the impact of tandem connectivity can be minimized, if the terminal secondary structure elements (SSEs) are flexible. In particular, two factors appear to facilitate TLD folding: (1) the interactions between terminal SSE are poorly ordered in the folding transition state, whereas nonterminal SSE are better structured, (2) the interactions between terminal SSE are weak in the native state. We apply these findings to wild-type proteins by examining experimental Φ-value data and by performing all-atom molecular dynamics simulations. We show that immunoglobulin-like domains appear to utilize the factors, which minimize the impact of tandem connectivity on their folding. Several single domain proteins, which are likely to misfold in tandems, are also identified. Proteins 2007;67:795–810.** © 2007 Wiley-Liss, Inc.

Key words: multidomain proteins; tandem-linked domains; immunoglobulin domains; protein folding and pathways; transition state ensemble

INTRODUCTION

Large proteins typically consist of individually folded domains covalently linked through their terminals into necklace-like tandems. However, experimental and theoretical studies of folding are usually focused on single domains (SD) with untethered terminals. The resulting folding mechanisms and pathways[1–3] are related to the assembly of native structure without the constraints imposed by the linkers to neighboring domains. Because individual domains in multidomain proteins unfold and refold being constrained by tandem connectivity, it is not a priori clear if the folding mechanisms of isolated domains are relevant to tandem-linked domains (TLD). An example of TLD folding and unfolding is given by mechanically active proteins, such as titin, fibronectin, tenascin, or filamins.[4–6] At least some of the domains in these proteins are repeatedly stretched and unfolded under physiological conditions.[7,8]

If domains in mechanically active proteins are forced to unfold, how do they fold back to their native states? Does folding of the domain linked in tandem differ from the folding of isolated single domain? Do tandem-linked domains utilize certain strategies to minimize the impact of tandem linkage? There are some indications that TLD and SD folding are different. Several experiments demonstrated that the folding of tandem-linked or terminal tethered domains is slower, sometimes significantly, than the folding of isolated single domains.[9–13] Recently, we examined folding of single domains initiated from temperature-denatured random coil and force-denatured stretched ensembles.[14] Compared to random coil initial ensemble the folding initiated with stretched conformations is slower and proceeds via modified pathway. The assembly of native interactions is delayed with respect to collapse, because domain must first contract and equilibrate in the random coil ensemble. Those simulations have demonstrated that even without tethering domain terminals changes in folding may be induced by using stretched initial conformations.[14]

The purpose of this study is to investigate general aspects of folding of multidomain tandems. Our computational strategy is based on the use of coarse grained models of proteins and construction of several heterogeneous tandems of domains. Our choice of tandem model is dictated by two considerations. First, to study the folding of TLD, reliable statistical data must be collected for

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.

*Correspondence to: Dmitri K. Klimov, Department of Bioinformatics, and Computational Biology, George Mason University, Manassas, VA 20110. E-mail: dklimov@gmu.edu

Received 30 September 2006; Revised 5 October 2006; Accepted 13 November 2006

Published online 22 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21339

large systems consisting of hundreds of amino acids. Because corresponding all-atom molecular simulations even with implicit solvent model are computationally costly, we use off-lattice coarse grained models. Second, we seek to identify generic basic consequences of tandem connectivity free of sequence specific effects. This objective is better served by coarse grained models, which retain only the most important characteristics of polypeptide chains, such as residue connectivity and heterogeneity and native topology.

By computing folding of tandem-linked domains and comparing it with the folding of isolated single domains, we provide some tentative answers to the questions posed earlier. We propose that tandem linkage of domains leads to two possible folding scenarios. In the first, tandem connectivity has a relatively minor impact on folding and the mechanisms of TLD and SD folding are similar. According to the second scenario, tandem linkage qualitatively changes folding mechanism. In particular, domains, which fold without intermediates via two-state mechanism as single domains, may misfold when inserted into tandems. The misfolded intermediates are created by topological constraints imposed by interdomain linkers. In both folding scenarios, tandem linkage considerably slows down folding. We also show that the impact of tandem connectivity is likely to depend on the flexibility of terminal secondary structure elements (SSEs). We conclude the article with the discussion on the applicability of our findings to wild-type proteins. Using reported Φ-value data and performing all-atom molecular dynamics simulations, we demonstrate that the folding of immunoglobulin-like domains in wild-type tandems is expected to be largely unaffected by domain connectivity. We also suggest several plausible examples of single domain proteins, which are likely to misfold in tandems.

METHODS

Construction of Multidomain Tandems

Because our goal is to identify generic consequences of tandem connectivity, we use off-lattice C_α-based coarse grained models of proteins, which are useful in exploring basic aspects of folding and aggregation.[3,15,16] To construct tandems we use four strand β-barrel domains S1 and S2 (Fig. 1) introduced previously.[17] S1 and S2 consist of $M = 46$ connected beads of three types, hydrophobic B, hydrophilic L, and neutral N. The sequences of S1 and S2 in three-letter code are $B_9N_3(LB)_3NBLN_2$ $LBN(BL)_4N_2LB_9$ and $B_9N_3(LB)_3NBLN_2B_9N_3(LB)_5L$, respectively. The potential energy E_p includes contributions from bond-length V_{BL} and bond-angle V_{BA} potentials, the dihedral angle potential V_{DIH}, and nonbonded potential V_{NB} described elsewhere.[18] The average separation between C_α-carbons along the sequence is $a = 3.8$ Å. The nonbonded interaction between a pair of hydrophobic residues B separated by the distance r is

$$V_{NB}^{BB}(r) = 4\lambda\epsilon_h\left[\left(\frac{a}{r}\right)^{12} - \left(\frac{a}{r}\right)^{6}\right], \quad (1)$$

where λ is a random factor accounting for the diversity of hydrophobic interactions and $\varepsilon_h = 1.25$ kcal/mol is the average strength of hydrophobic interactions, which serves as an energy unit in the model. The interactions between all other residues are repulsive.[18] The β-barrel fold is stabilized by proper distribution of hydrophobic residues and λ factors. The energy function includes native and nonnative attractive interactions, but does not distinguish chiral states. The native structures of S1 and S2 have 106 contacts (defined with 6.8 Å cut-off) and the potential energies of −85.5 and −88.0 kcal/mol, respectively.

There are important differences in S1 and S2 native structures. In S1 [Fig. 1(a)], the interactions between N- and C-terminal strands β1 and β4 are about 50% stronger than the interactions between any other β-strands. Middle strands β2 and β3 form relatively few native contacts. Structural fluctuations in the native state can be evaluated by computing standard deviation in the distance between strands βi and βj at the temperature of simulations, $\delta R(\beta i, \beta j)$ [Eq. (4)]. In S1, the β1–β4 interface is the most rigid ($\delta R(\beta 1, \beta 4) \approx 0.65$ Å), whereas the fluctuations in other interstrand distances are at least 60% larger.

In S2, the first three strands form a tightly packed stable core β1–β2–β3, whereas the terminal β4 is engaged in relatively few interactions with the core [Fig. 1(b)]. As a result, the most rigid part of the S2 native state is the β1–β2–β3 core, in which the interstrand fluctuations ($\delta R(\beta 1,\beta 2) \approx \delta R(\beta 1,\beta 3) \approx 0.7$ Å) are about four times smaller than the fluctuations in the distances between β4 and other strands. Therefore, in S1 the interface between terminal β-strands is rigid, whereas it is highly flexible in S2. These properties are relevant for the folding kinetics of these domains in multidomain tandems.

The tandems S2-S1-S2 and S2-S2-S2 are constructed by linking "head-to-tail" S1 and S2 domains using flexible linkers of five neutral N residues [Fig. 1(a,b)]. Interdomain interactions are limited to steric repulsion and the linkers are sufficiently long to minimize perturbations in domains' native states. S2-S1-S2 and S2-S2-S2 tandems are designed to study the TLD folding of the "middle" S1 and S2 domains, in which both terminals are constrained by linkers. We have tested two-domain constructs, but found that their folding displays no qualitative changes compared to SD folding. This outcome is due by the fact that only one domain's terminal in a dimer is constrained by linkers. Our results reported here show that a three domain tandem is a minimal system, which may show qualitative impact of tandem connectivity.

Simulation Details for Coarse-grained Model Tandems

To test the stability of the tandems we performed simulated annealing Langevin dynamics simulations.[18] We found the equilibrium properties of linked and isolated domains to be similar. The folding transition tem-

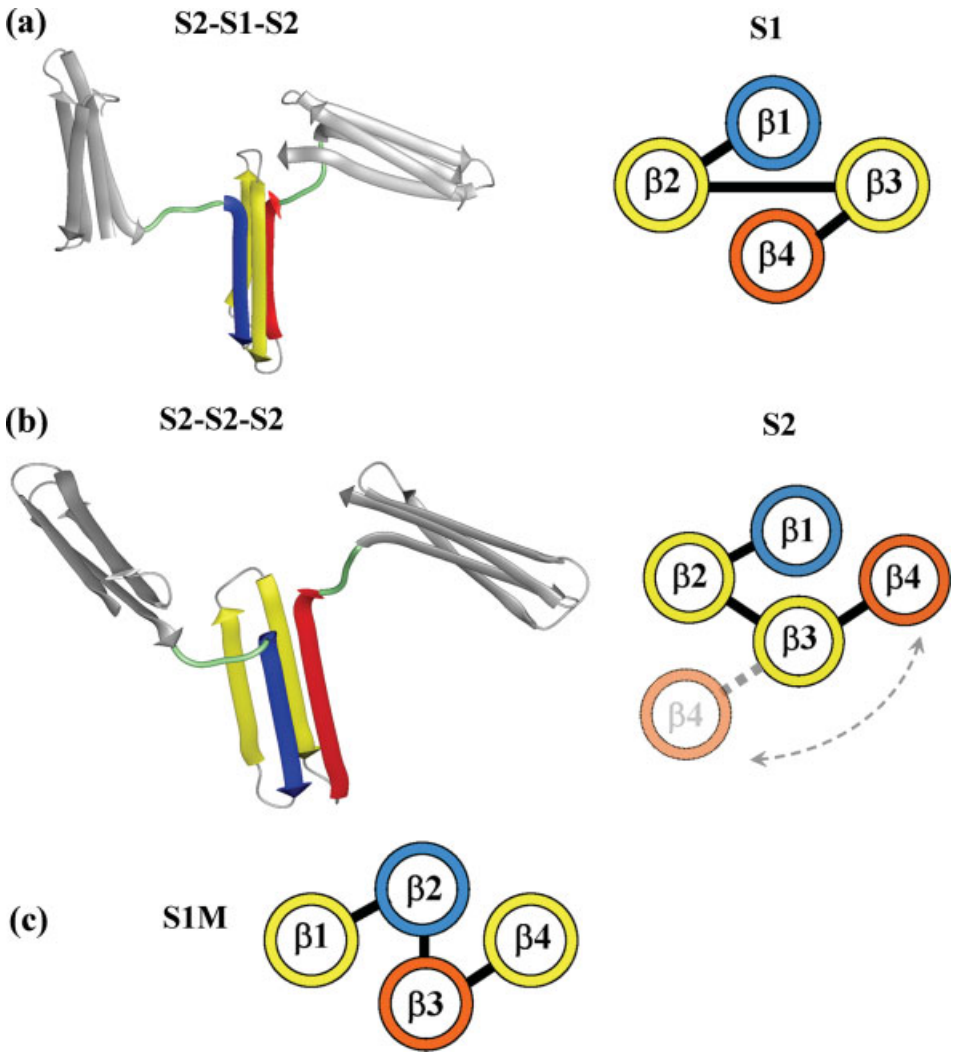

Fig. 1. (**a**) The tandem S2-S1-S2 and the native topology of the middle (in color) S1 domain. The native core of S1 is formed by the terminal strands β1 (in blue) and β4 (in red). The β1–β4 interface is highly stable and rigid. (**b**) The tandem S2-S2-S2 and the native topology of S2 domain. In contrast to S1, the rigid native core of S2 includes nonterminal strands β2 and β3 and excludes the terminal strand β4, which has low stability and fluctuates between native and nonnative (shown by dashed line) positions. Nonnative position of β4 is associated with the cluster CL2 transiently populated in S2 TLD folding (Table I). (**c**) The native topology of S1M domain, which is derived from S1 by reconnecting β-strands. In S1M β2 and β3 (formerly β1 and β4 in S1) are covalently linked and form a tight native core, whereas the interactions between the terminal strands are weak.

peratures for S1 and S2, T_F, approximately coincide. Therefore, TLD simulations may be performed under the condition of isostability for all domains in a tandem.

The TLD folding kinetics was obtained using Langevin dynamics at water viscosity. Folding is studied at the temperature $T_S < T_F$, for which the equilibrium fraction of native contacts in all domains is $\langle Q(T_s)\rangle \approx 0.7$. Refolding kinetics is monitored starting with either stretched (by 90% of the contour tandem length) or random coil conformations. To keep the error in folding timescales within 10%, 100–300 independent trajectories were generated. To monitor the approach to a native state we used domain overlap function[19]

$$\chi(t) = 1 - \frac{2}{M^2 - 5M + 6}\sum_{i=1}^{M-3}\sum_{j=i+3}^{M}\Theta(\Delta - |r_{ij}(t) - r_{ij}^0|), \quad (2)$$

where r_{ij} and r_{ij}^0 are the distances between the residues i and j in the current (at a time t) and native conformations, $\Delta = 0.2a$, and $\Theta(x)$ is a Heaviside step function. According to Eq. (2), $\chi = 1$ and 0 correspond to fully unfolded and native states, respectively. The first passage time to the native state τ_{li} in a trajectory i corresponds to the first instance, when $\chi(t) < \chi_0 = 0.15$. Given the distribution of τ_{li}, the fraction of unfolded molecules $P_u(t)$ is computed, from which the folding time

$\tau_F = \int_0^\infty P_u(t)\, dt$.[18] Folding of a domain is characterized by the fraction of native contacts Q, the fraction of native contacts formed by a strand βs, $Q_{\beta s}$, and the radius of gyration R_g. These probes were averaged over hundreds of folding trajectories (indicated by brackets $\langle \ldots \rangle$) and fit using exponentials to extract timescales and amplitudes of kinetic phases.

Molecular Dynamics Simulation

All-atom explicit solvent molecular dynamics (MD) simulations probe the energetics of native structures of wild-type domains. The CHARMM22 force field and NAMD program[20] were used. To reduce computational costs domain's native structure was solvated in a water sphere, which extends at least 12 Å from any protein atom. To maintain water density and sphere shape spherical boundary conditions were applied to water only. A translational drift of a protein is controlled by coupling protein's center of mass to the sphere center with harmonic potential. Production trajectories were obtained using Langevin dynamics with the damping coefficient of 5 ps^{-1}. The temperature of production simulations is set equal to the temperature of relevant folding experiments (e.g. 298 K for I27[21]).

The fluctuations in the position of a residue i, $\vec{R}_i(t)$, were evaluated using the root mean-squared displacement (RMSD) δR_i given by

$$\delta R_i = \left[\frac{1}{(\tau - \tau_{eq})} \int_{\tau_{eq}}^{\tau} (\vec{R}_i(t) - \vec{R}_i(0))^2 \, dt \right]^{1/2}, \qquad (3)$$

where τ and τ_{eq} are simulation and equilibration times, $\vec{R}_i(0)$ is a native position. Positions of residues are given by the coordinates of C_α-carbons. In computing Eq. (3), protein conformations were first adjusted to minimize the total root mean-squared deviation δR with respect to the native structure. The length of simulations τ is determined from the requirement that MD trajectory reaches quasi-equilibrium in a native energy basin evidenced by a flat region in $\delta R(t)$ at $t > \tau_{eq}$. For I27 τ_{eq} = 2 ns and τ = 6 ns (see Supplementary Materials). The average δR computed for $t > \tau_{eq}$ is 1.5 Å, but is reduced to 0.9 Å if turn regions are excluded. Similar results were obtained for other wild-type domains.

The mobilities of native SSE were evaluated using the standard deviations $\delta R(s_1, s_2)$

$$\delta R(s_1, s_2) = \left[\frac{1}{(\tau - \tau_{eq})} \int_{\tau_{eq}}^{\tau} (R(s_1, s_2, t) - \overline{R(s_1, s_2)})^2 \, dt \right]^{1/2}, \qquad (4)$$

where $R(s_1, s_2, t)$ is the distance between the centers of mass of the SSE s_1 and s_2 at a time t and horizontal bar implies averaging over $\tau - \tau_{eq}$.

Progress Variable Cluster Method

To compute folding transition state ensemble (TSE) we use progress variable cluster method.[22] In short, structures sampled in a folding trajectory were grouped into clusters using pattern recognition algorithm.[22,23] To filter out stochastic fluctuations a trajectory is represented by a sequence of clusters. In each trajectory i, we monitored the formation of stable native contacts as a function of the progress variable δ defined as $\delta = t/\tau_{1i}$, where t is time. A contact is considered stable at δ, if it remains formed (with some tolerance for short-lived disruptions) until δ = 1. A sharp increase in the fraction of stable native interactions $P(\delta)$ (or its derivative $dP/d\delta$) is attributed to the passage through TSE at δ_{TSE}. In our previous studies,[22] we confirmed that progress variable cluster method and "Pfold" (stochastic separatrix) method[24] yield consistent results.

RESULTS

Folding of Tandem-linked Domains

Folding of single and tandem-linked domains (SD and TLD, respectively) is studied using single domains S1 and S2 and three-domain tandems S2-S1-S2 and S2-S2-S2 (see Methods and Fig. 1). The selection of S1 and S2 is motivated by the differences in their native energetics, which allow them to qualitatively represent diverse wild-type domains. The SD folding of S1 has been reported.[14] The time scale of SD collapse obtained from fitting the radius of gyration $\langle R_g(t) \rangle$ is τ_c^{SD} = 84 ns, while the time scale of forming native interactions extracted from the fit of the fraction of native contacts $\langle Q(t) \rangle$ is τ_Q^{SD} = 87 ns. Folding of individual β-strands takes place synchronously and approximately coincide with τ_Q^{SD}. The folding time scale obtained from fitting the fraction of unfolded molecules $P_u(t)$ is τ_F = 127 ns [Fig. 2(a)]. Because τ_F registers simultaneous formation of almost all native interactions (see Methods), $\tau_F > \tau_Q$. Thus, SD folding of S1 is cooperative and approximately two-state with all elements of native structure forming simultaneously. The SD folding mechanism of S2 differs from that of S1, because native interactions associated with β-strands form in two stages [Fig. 3(a)]. During the first, three strands β1, β2, β3, which form the native core (see Methods), cooperatively fold with the timescales $\tau_{Q\beta1}^{SD} \approx \tau_{Q\beta2}^{SD} \approx \tau_{Q\beta3}^{SD} = 35$ ns. The second stage occurs later and is due to the docking of β4 ($\tau_{Q\beta4}^{SD} = 55$ ns). The SD folding time is τ_F = 89 ns [Fig. 2(b)]. The collapse time of S2 (τ_c^{SD} = 46 ns) approximately coincides with the folding of β-strands.

The S2 TLD folding is studied by monitoring the folding of the "middle" domain in S2-S2-S2 tandem. To imitate refolding in wild-type tandems after mechanical stretching, TLD folding was initiated with stretched conformations [Fig. 3(b)]. Using two exponential fit, the relaxation of $\langle R_g(t) \rangle$ is described by the fast collapse timescale τ_{c1}^{TLD} = 3 ns and the slow collapse timescale τ_{c2}^{TLD} = 277 ns [Fig. 3(b)]. Acquisition of native interac-

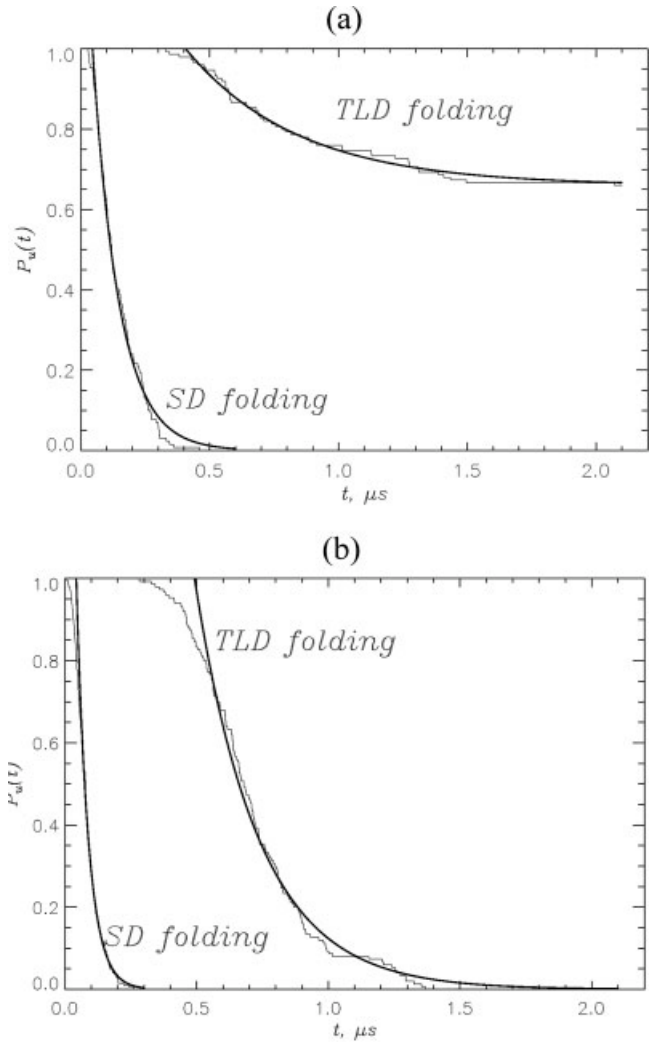

Fig. 2. The fractions of unfolded domains S1 (**a**) and S2 (**b**) $P_u(t)$ as a function of time (thin line). TLD folding implies folding of the middle S1 (or S2) domain in the tandem S2-S1-S2 (or S2-S2-S2). Although the TLD folding time of S2 increases almost ten-fold, its folding mechanism is similar to that observed in SD folding (Fig. 3). In contrast, TLD folding of S1 partitions into folded and misfolded phases due to formation of the intermediate **I**, which is not observed in SD folding. Black thick curves represent single exponential fits. For S1 TLD folding the fit characterizes the folded kinetic phase.

tions remains single exponential characterized by the timescale $\tau_Q^{TLD} = 518$ ns. Similar to SD folding ordering of the β-strands occurs in two stages. Figure 3(b) shows that the native three-strand core is cooperatively formed ($\tau_{Q\beta1}^{TLD} = 490$ ns, $\tau_{Q\beta2}^{TLD} = 481$ ns, $\tau_{Q\beta3}^{TLD} = 506$ ns), while folding of the strand β4 occurs later ($\tau_{Q\beta4}^{TLD} = 664$ ns). The total TLD folding time is $\tau_F^{TLD} = 735$ ns.

Compared to SD folding, TLD folding pathway reveals an increase in folding times and separation of folding from collapse. It follows from Figure 3(b) that by the time $\langle R_g(t)\rangle$ decreases in half of its initial value, the native content measured by $\langle Q_s(t)\rangle$ (s = β1, …, β4) increases by less than 5%. Consequently, the collapse time represented by τ_{c2}^{TLD} is almost twice shorter than τ_Q^{TLD}. However, apart from the differences at the early stages of folding and distinct timescales, TLD and SD pathways remains similar. Both feature two stages in the assembly of native conformation, delayed folding of β4, and no folding intermediates are observed. Therefore, the main consequence of linking S2 domain in a tandem is a sharp increase in folding time.

TABLE I. Characteristics of Domain Structures in TLD Folding[a]

Cluster	Q	$Q_{\beta1\beta2}$	$Q_{\beta1\beta3}$	$Q_{\beta1\beta4}$	$Q_{\beta2\beta3}$	$Q_{\beta2\beta4}$	$Q_{\beta3\beta4}$	R_g/R_g^{0}[b]
Middle S1 domain in S2-S1-S2 tandem								
CLN	0.9	1.0	1.0	1.0	—	1.0	1.0	1.0
CLI1	0.7	0.9	0.7	0.9	—	0.4	0.5	1.0
CLI2	0.8	0.4	0.8	0.9	—	0.9	0.8	1.0
Middle S2 domain in S2-S2-S2 tandem								
CLN	0.9	1.0	0.9	0.9	0.9	—	0.9	1.0
CL1	0.6	0.8	0.9	0.4	0.6	—	0.7	1.0
CL2	0.6	0.8	0.9	0.2	0.8	—	0.3	1.1

[a]The final conformations of middle domains obtained from TLD folding trajectories for S2-S1-S2 (150 structures) and S2-S2-S2 (100 structures) tandems. Only clusters, incorporating at least 10% of structures, are listed.
[b]Subscript "0" implies native conformation.

Is the result that TLD and SD folding pathways are similar general? To answer this question, we examined the TLD folding of the "middle" domain S1 in the tandem S2-S1-S2 starting with stretched conformations. Similar to S2 TLD folding, two exponential fit of $\langle R_g(t)\rangle$ yields two time scales in S1 collapse, corresponding to fast contraction at $\tau_{c1}^{TLD} = 5$ ns and slower decrease in S1 dimensions at $\tau_{c2}^{TLD} = 488$ ns (data not shown). The distinctive feature of S1 TLD folding is that S1 reaches the native state in only 34% of 2.1 μs trajectories [Fig. 2(a)]. Importantly, in the last quarter of TLD simulation time, very few folding events occur. These observations suggest a formation of misfolded intermediate **I**, which blocks folding in significant fraction of S1 domains. The result requires further investigation, because as a single domain S1 folds without populating intermediates. The nature of **I** is described in detail in the next section.

To further analyze folding mechanism of S1, we partitioned TLD trajectories into folded and misfolded. In the folded phase, all β-strands fold cooperatively on a single time scale coinciding with the time scale of the assembly of native fold (for the folded phase $\tau_Q^{TLD} = 891$ ns). Therefore, if S1 avoids misfolding, its TLD folding qualitatively resembles the SD folding.

TLD Folding Intermediate I

To probe the origin of **I**, we selected the final S1 conformations in TLD trajectories and performed steepest descent simulations to map local energy minima. Clustering[22] of the energy minimized conformations yields three structurally distinct clusters. Two clusters, CLI1 and CLI2 (Table I and Fig. 4), which appear only in misfolded trajectories, represent the intermediate **I**. The distinct feature of **I** is a nonnative arrangement of β-

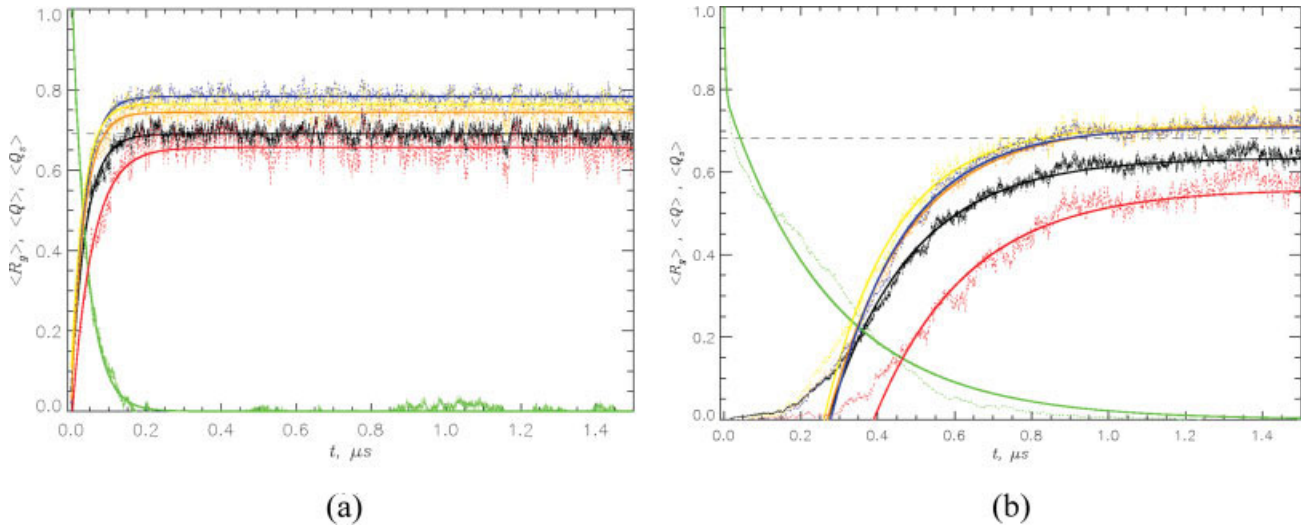

Fig. 3. SD (**a**) and TLD (**b**) folding of S2 is monitored by the radius of gyration, $\langle R_g(t)\rangle$ (in green), the fraction of native contacts, $\langle Q(t)\rangle$ (in black), and by the fractions of native contacts formed by β-strands, $\langle Q_s(t)\rangle$ ($s = \beta1,\ldots,\beta4$). The color codes for the strands are β1 (blue), β2 (yellow), β3 (orange), and β4 (red). Although an increase in TLD folding times and separation of TLD folding and collapse are observed, the SD and TLD folding pathways remain similar, because the order of folding of β-strands in both panels is identical. All quantities are normalized to vary from 0 to 1. The thick curves are exponential (biexponential for $\langle R_g(t)\rangle$ in (b)) fits. Averaging is done over 100 trajectories. The dashed lines indicate equilibrium values of Q.

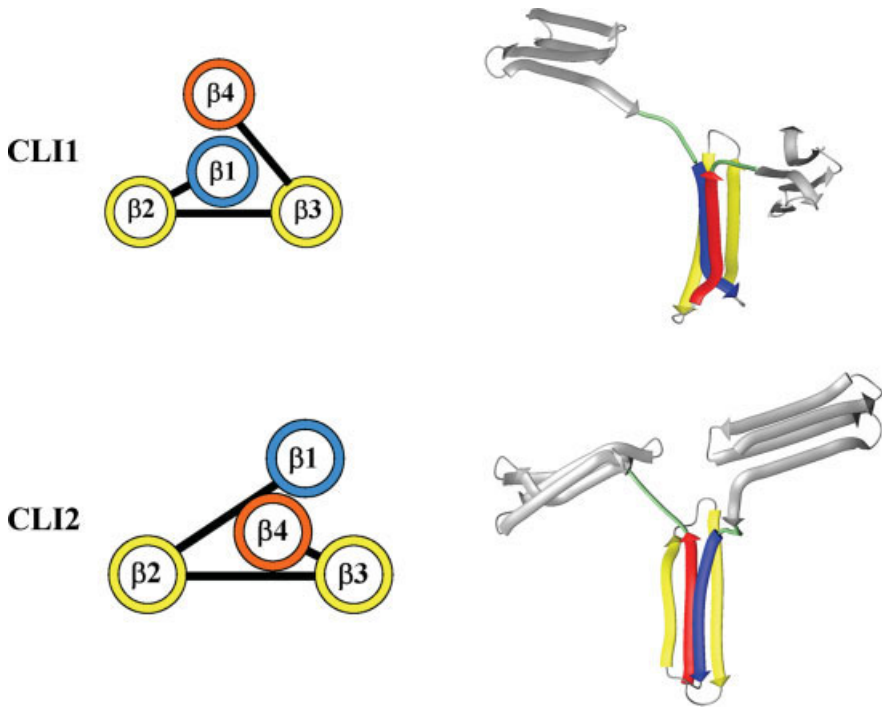

Fig. 4. The topologies and representative structures from two conformational clusters, CLI1 and CLI2, which constitute the S1 TLD folding intermediate **I**. CLI1 and CLI2 differ with respect to the arrangement of the terminal β-strands β1 and β4. The probabilities of occurrence of CLI1 and CLI2 in TLD folding are 0.47 and 0.14, respectively. The native cluster CLN [Fig. 1(a)] is sampled only in the folded TLD phase (the probability of occurrence is 0.34). Conversion of CLI1 and CLI2 into CLN is blocked by the linkers (in green) to neighboring domains (in grey).

strands, in which both β1 and β4 are positioned on the same side of the plane defined by the middle β-strands (β2–β3). In the native cluster CLN, β1 and β4 are positioned on the opposite sides of (β2–β3) [Fig. 1(a) and Table I]. CLI1 and CLI2 are composed of low energy conformations with significant native content (their average potential energy −68 kcal/mol is close to the native energy of −85.5 kcal/mol).

Figure 4 shows that CLI1 and CLI2 differ with respect to the positions of terminal strands β1 and β4. This figure also reveals the factors impeding the rearrangement of **I** into the native state **N**. CLI1 and CLI2 can reach **N** by transient unfolding of either β4 (CLI1) or β1 (CLI2) and refolding these strands on the opposite side of (β2, β3) (pathway 1). The second pathway from CLI1 and CLI2 to **N** involves relocating β1 (CLI1) or β4 (CLI2) to the opposite side of (β2, β3) through the cleft between β2 and β3. The first **I**→**N** pathway does not conflict with the interdomain linkers, but is energetically unfavorable because of the strong native β1–β4 interactions present in **I** (Table I). Figure 5(a) shows that after formation of β1–β4 interface it remains highly stable and transient unfolding of the terminal strands does not occur. The second **I**→**N** pathway, which although does not break β1–β4 interactions, is topologically blocked by the interdomain linkers (Fig. 4). If the linkers are removed, **I** readily converts into **N** within 31 ns $\approx 0.25\tau_F^{SD}$.

Folding TSE and Pathways in Multidomain Tandems

We first study the transition state ensemble (TSE) and pathways for TLD and SD folding of S2. Our goal is to compare nucleation mechanism in single and tandem-linked domains using a progress variable cluster method[22] (see Methods). Figure 6 shows the dependence of the fractions of stable (nucleation) native contacts formed by the β-strand β*s*, $P_{\beta s}^{SD}$ (δ), in SD folding (dashed lines). The propagation of stable native contacts shows that those coupling β4 with other strands are established significantly later than the stable native interactions between other β-strands. TSE region for S2 is crossed at $\delta_{TS}^{SD} = 0.93$, which corresponds to a sharp growth in $dP^{SD}/d\delta$ (see Methods). We also computed the probability $P_n(\delta)$, which gives the fraction of stable native interactions formed by the residue n at δ. The inset to Figure 6 shows the distributions of P_n^{SD} (δ) at and around δ_{TSE}^{SD} for SD folding. Although the plot generally demonstrates a delocalized nature of nuclei,[26,27] nucleation contacts tend to concentrate in the three-strand native core β1–β2–β3. Only about one-third of native β1–β4 contacts are present in the TSE. These findings are in agreement with the analysis of SD folding discussed earlier [Fig. 3(a)].

The fractions of stable native contacts formed by β*s* in TLD folding, $P_{\beta s}^{TLD}$ (δ), are displayed in Figure 6, in which the inset shows the distribution of the probabilities P_n^{TLD} (δ) for individual residues. We determined that for TLD folding $\delta_{TSE}^{TLD} \approx 0.9825$ (see Methods). Comparison of the distributions of $P_n^{TLD}(\delta_{TSE})$ and $P_n^{SD}(\delta_{TSE})$ indicates that S2 TSE remains virtually unchanged upon incorporation of this domain into tandem. The most stable nucleation contacts are located in β2 and also β1 and β3 strands. The strand β4 is largely unstructured and the fraction of native interactions formed in the β1–β4 interface is just 0.2. Therefore, the SD and TLD folding in S2 is largely initiated with the middle strands. Note that TLD folding TSE is positioned closer to the native state **N** than SD TSE. Taking into account vastly different SD and TLD folding time scales, we find that the SD and TLD time intervals between passing through TSE and reaching **N** are comparable (≈ 6 and 13 ns, respectively). Therefore, movement of TSE toward **N** in TLD folding reflects a longer search for folding nuclei in the domains linked in tandems. Once TSE is reached, the descent to **N** is not significantly affected by the linkage. It is important to point out that the movement of TSE is not related to Hammond postulate, because δ progress variable is not a structural reaction coordinate.

The TSE analysis for S1 TLD folding can only be done for those trajectories, which avoid misfolding. Although such analysis would not reflect the full impact of domain connectivity, it is still instructive to compare the TSE in SD and TLD (folded phase) trajectories. In SD folding, nucleation contacts between terminal β1 and β4 are developed prior to the formation of nucleation interactions associated with the middle strands β2 and β3. As a result, the most structured regions in the SD TSE are in the terminal β-strands. For example, the fraction of native β1–β4 interactions formed in the TSE is 0.9, i.e., β1–β4 interface is almost completely structured. The formation of nuclei and the structure of TSE in TLD (folded phase) folding are qualitatively similar to those observed in SD folding. Therefore, if S1 avoids the kinetic trap created by interdomain linkers, its TLD folding follows the same mechanism as in SD folding.

TLD Folding and Native Energetics

To establish a connection between TLD folding scenarios and the energetics of native states, we compare S1 and S2 native states and perform targeted mutations. As described in Methods, the native structure of S1 has two distinct features: (i) highly stable β1–β4 interface, (ii) virtually no interactions between the strands β2 and β3. These characteristics are responsible for the misfolding of S1 in TLD folding. To illustrate this, we created a mutant S1M by reconnecting the β-strands as shown in Figure 1(c). In S1M most stable native interactions occur between the middle β-strands β2 and β3, whereas there are almost no interactions between terminal strands. In TLD simulations of S2-S1M-S2 tandem S1M folds on a time scale of $\tau_F^{TLD} = 819$ ns with no instances of misfolding or formation of the intermediate **I**.

In contrast to S1, TLD folding of S2 remains robust. In its native structure the strongest attractive interactions occur between the strands β1 and β3, whereas the β1–β4 interactions are weak (see Methods). The analysis of S2-S2-S2 TLD folding trajectories shows that misfolded structures [such as shown in Fig. 1(b) by dashed line] do transiently form, but are rapidly converted to **N**. Consistently, the energy minimized conformations of the middle S2 from the S2-S2-S2 tandem can be grouped into three clusters (Table I). The native cluster CLN occurs with the probability 0.45 and has native-like arrangement of β strands. The cluster CL1 [probability

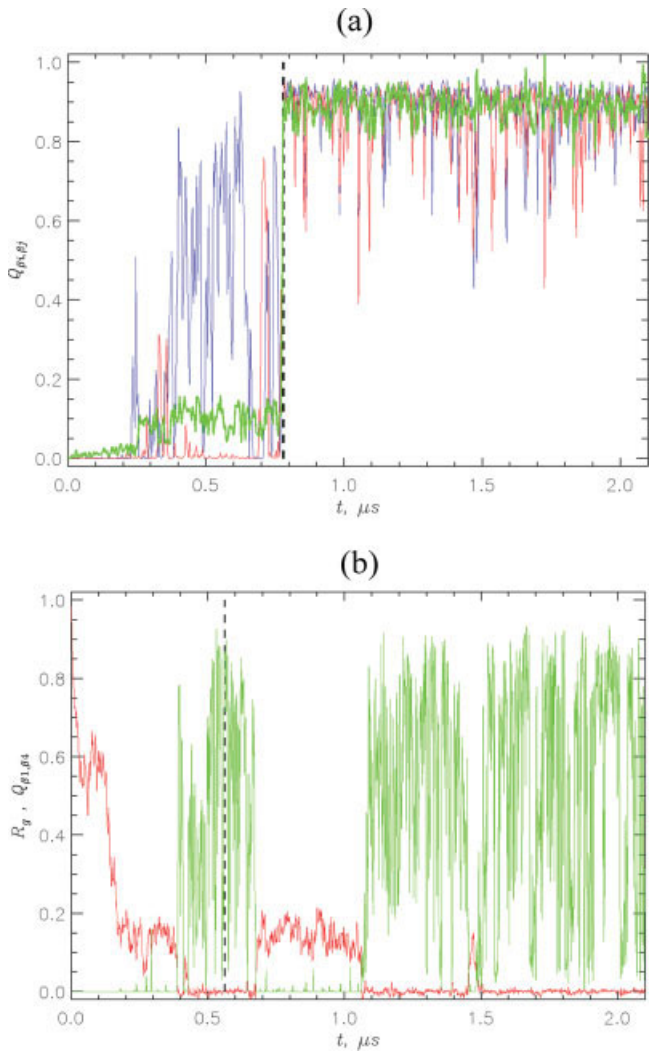

Fig. 5. (**a**) Formation of interstrand interactions in S1 in a typical folded TLD trajectory for the S2-S1-S2 tandem. The fractions of native contacts between β-strands are colored as $Q_{\beta1,\beta2}(t)$ (in blue), $Q_{\beta3,\beta4}(t)$ (in red), $Q_{\beta1,\beta4}(t)$ (in green). Once formed β1–β4 interface experiences little fluctuations and prevents the escape from the intermediate **I**. (**b**) The fraction of native interactions between the terminal strands $Q_{\beta1,\beta4}(t)$, in the middle S2 of the S2-S2-S2 tandem for a typical TLD trajectory. In contrast to S1, there are frequent fluctuations in β1–β4 interactions even after complete assembly of the S2 native fold. The time dependence of the radius of gyration $R_g(t)$ is shown in red. In both panels, dashed vertical lines mark the first passage time to the native state.

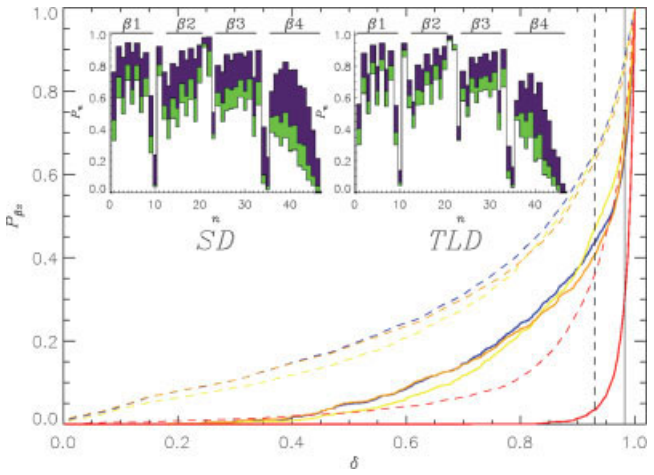

Fig. 6. Formation of folding nuclei in S2 is probed by the fractions of stable native contacts formed by strand β*s* (s = 1,. . .,4), $P_{\beta s}$, as a function of progress variable δ. $P_{\beta s}(\delta)$ for SD and TLD folding are represented by dashed and solid curves, respectively. The color codes for β-strands are the same as in Figure 3. The vertical dashed and solid lines indicate the location of transition states for SD (δ_{TSE}^{SD}= 0.93) and TLD (δ_{TSE}^{TLD} = 0.9825) folding, respectively. The insets show the fractions of stable native contacts P_n formed by the residue *n* at and near crossing the TSE for SD and TLD folding. The $P_n(\delta_{TSE})$ profile is given in green, those in white and purple are obtained at δ = 0.88 and 0.98 for SD folding and at δ = 0.9725 and 0.9925 for TLD folding. S2 β-strands are shown on top. Formation of nuclei and TSE are similar in SD and TLD folding. Therefore, S2 folding is not significantly affected by tandem connectivity. The distributions of $P_n(\delta)$ are analogous to experimental Φ-value distributions.[25]

of occurrence 0.31, shown in Fig. 1(b) by dashed line] has a lower native content and nonnative arrangement of β4. The cluster CL2 (the probability of occurrence 0.24) is partially unfolded. Interestingly, the times of TLD folding from CL1 and CL2 are $0.09\tau_F^{TLD}$ and $0.25\tau_F^{TLD}$, respectively. Misfolded S2 structures within the tandem rapidly reach **N** because of the flickering of β1–β4 native interactions [Fig. 5(b)].

Using S2 as an example, we can assume that the TLD folding is facilitated by the formation of stable native core made of "middle" strands and by the flexibility (low stability) of terminal strands (e.g. β4). If so, S2 can be forced to misfold in a tandem according to the following recipe: β2–β3 interactions must be weakened, but the β1–β4 native contacts must be strengthened. To verify this assumption we created S2 mutant, S2M, in which all contact energies between β4 strand and other strands were strengthened by 4 kcal/mol, whereas the β2–β3 contact energies were destabilized by 2 kcal/mol. As a result, the mutant S2M incorporates two energetic "flaws" of S1 and folds on a dramatically longer time scale of τ_F^{TLD} = 4.3 μs. Within the length of simulations (2.1 μs) the native yield is 63%. Therefore, incorporation of the S1-like features forces S2 to misfold in tandems.

TLD Folding Initiated with Random Coil States

Using the tandems S2-S1-S2 and S2-S2-S2, we have explored the TLD folding of the middle S1 and S2 domains starting with random coil states. The folding and collapse of S2 occur two- to three-times slower than in SD folding. Interestingly, as in SD folding the collapse and folding are synchronous and folding of the strand β4 is delayed compared to the folding of β1, β2, or β3. Therefore, apart from some increase in folding times, there are no appreciable differences in the SD and TLD folding initiated with random coils. The TLD folding of S1 qualitatively differs from its SD folding. Folding is completed (within 2.1 μs) in just 34% of trajectories,

while two-thirds of them become trapped in the stable intermediate $\mathbf{I}'$. Structural analysis confirms that $\mathbf{I}'$ is identical to $\mathbf{I}$ observed in the S1 TLD folding initiated with stretched conformations.

Therefore, two effects of tandem linkage are independent on folding initial conditions. These are an increase in time scales for folding and formation of new TLD intermediates in S1. Also, irrespective of initial conditions, the TLD folding pathway of S2 is similar to that in SD folding. Stretched initial conditions do however introduce an element, which is not seen in TLD folding initiated with random coil states, that is the separation of collapse and folding.[14]

Topological Effects are the Main Factors in TLD Folding

In this section, we investigate the physical factors responsible for the increase in TLD folding times and misfolding. To this end, we attached two identical large repulsive spheres O of the radius $R_O = 40$ Å $\approx 10.5a$ to S1 terminals, creating a O-S1-O system. Following Stokes formula the friction coefficient for O was set to $\zeta_O = 10\zeta$, where ζ is a friction coefficient applied to amino acids. The attachment of the spheres, which mimic neighboring domains, is motivated by the observation that in 97% of TLD trajectories for S2-S1-S2 tandem S1 folds (or forms $\mathbf{I}$) after folding of both S2. (Similar results were obtained for the middle S2 in S2-S2-S2 tandem.) Because the bond O-S1 has no own excluded volume, it can be volume crossed by a protein chain. The crossing of O-S1 bond will not occur, if the bond length l is sufficiently small, i.e. $l \approx R_O + a$. In this case, steric repulsions of S1 and O would repel the polypeptide chain from crossing the O-S1 bond. Therefore, if we vary l keeping R_O constant, we would change the strength of steric repulsion between the bond (which represents interdomain linkers) and domain. Specifically, we consider four values of l: $l_1 = 15a$, $l_2 = 14.1a$, $l_3 = 13.7a$, $l_4 = 13.3a$. Initiating folding with stretched conformations, we found that the folding time strongly depends on l, increasing from 0.67 μs (l_1) to 7.2 μs ($l = l_4$). Simultaneously, the fraction of molecules folded within the simulation time (2.1 μs) decreases from 1.0 to 0.22. The final conformations in the trajectories, which failed to fold, are similar to the intermediate $\mathbf{I}$ observed in TLD folding of S1. The simulations using O-S2-O system do not result in misfolding, but register significant increase in folding times for the short O-S2 linkers. Therefore, attached spheres emulate the effects of tandem linkage on domain folding. It appears that the main factor affecting TLD folding is the excluded volume of the linkers and not of the domains.

To investigate the contribution of solvent friction in constraining the motions of TLD terminals, we considered the system O-S2-O, in which the length of the linker bond was set to $l = 13.7a$. The folding simulations were done using $R_O = 10.5a$ and $\zeta_O = 10\zeta$. The folding time for O-S2-O is $\tau_F = 662$ ns. We then create two versions of O-S2-O. In the first, $(\text{O-S2-O})_1$ the friction coefficient for O was reduced to $\zeta_O = \zeta$ keeping R_O constant. In the second version, $(\text{O-S2-O})_2$, the radius of the sphere is reduced to $R_O = 1.0a$ with ζ_O fixed. $(\text{O-S2-O})_2$ folds on a time scale $\tau_F = 210$ ns that is more than three times faster than O-S2-O, while the folding time for $(\text{O-S2-O})_1$ ($\tau_F = 449$ ns) is reduced only by 30%. To clarify if the decrease in τ_F for $(\text{O-S2-O})_2$ is the consequence of reduced steric effect created by the linkers (bonds) or spheres, we considered the system $(\text{O-S2-O})_3$, which differs from O-S2-O only by increased length of the O-S2 bond ($l = 20a$). We found that the folding time for $(\text{O-S2-O})_3$ $\tau_F = 209$ ns coincides with τ_F for $(\text{O-S2-O})_2$. Therefore, topological effect created by interdomain linkers appears to be the dominant factor in TLD folding.

DISCUSSION

Consequences of Tandem Connectivity

Simulations of model multidomain tandems suggest two possible consequences of domain connectivity (Fig. 7). Scenario 1 corresponds to the case, when the linkage of domains (such as S2) does not alter their folding significantly. In such tandems TLD folding is slower than SD folding, especially when stretched initial conditions are used. In this case, τ_F may increase by, at least, an order of magnitude [Fig. 2(b)]. Stretched initial conditions also lead to a separation between collapse and folding [Fig. 3(b)]. Apart from these relatively modest differences, the folding pathway and transition state ensemble remain unaltered (Fig. 6). Scenario 2 in Figure 7 is applied to the domains, which experience misfolding within tandems (such as S1). Their two-state SD folding is qualitatively different from the folding in tandems. TLD folding, initiated from either stretched or random coil states, leads to partitioning of molecules into folded and misfolded phases similar to kinetic partitioning observed in slow folding proteins.[28–30] Misfolded intermediates occur, when domain's terminals constrained by linkers cannot rearrange into native positions. It is important to note that due to limitations of the model the proposed scenarios are unlikely to account for all the consequences of tandem connectivity. This point is further highlighted in the Conclusions.

Factors Facilitating TLD Folding

On the basis of our simulations, we propose that the important requirement for efficient TLD folding is a conformational flexibility of the terminal secondary structure elements (SSE). From the analysis of S1 and S2 native energetics, folding pathways, and TSE, we identify two specific factors, which facilitate TLD folding. These are:

(1) the interactions between terminal SSE are poorly ordered (or absent) in the TSE, whereas nonterminal SSE are better structured;
(2) in the native state the interactions between terminal SSE are weak and their interface experiences large conformational fluctuations.

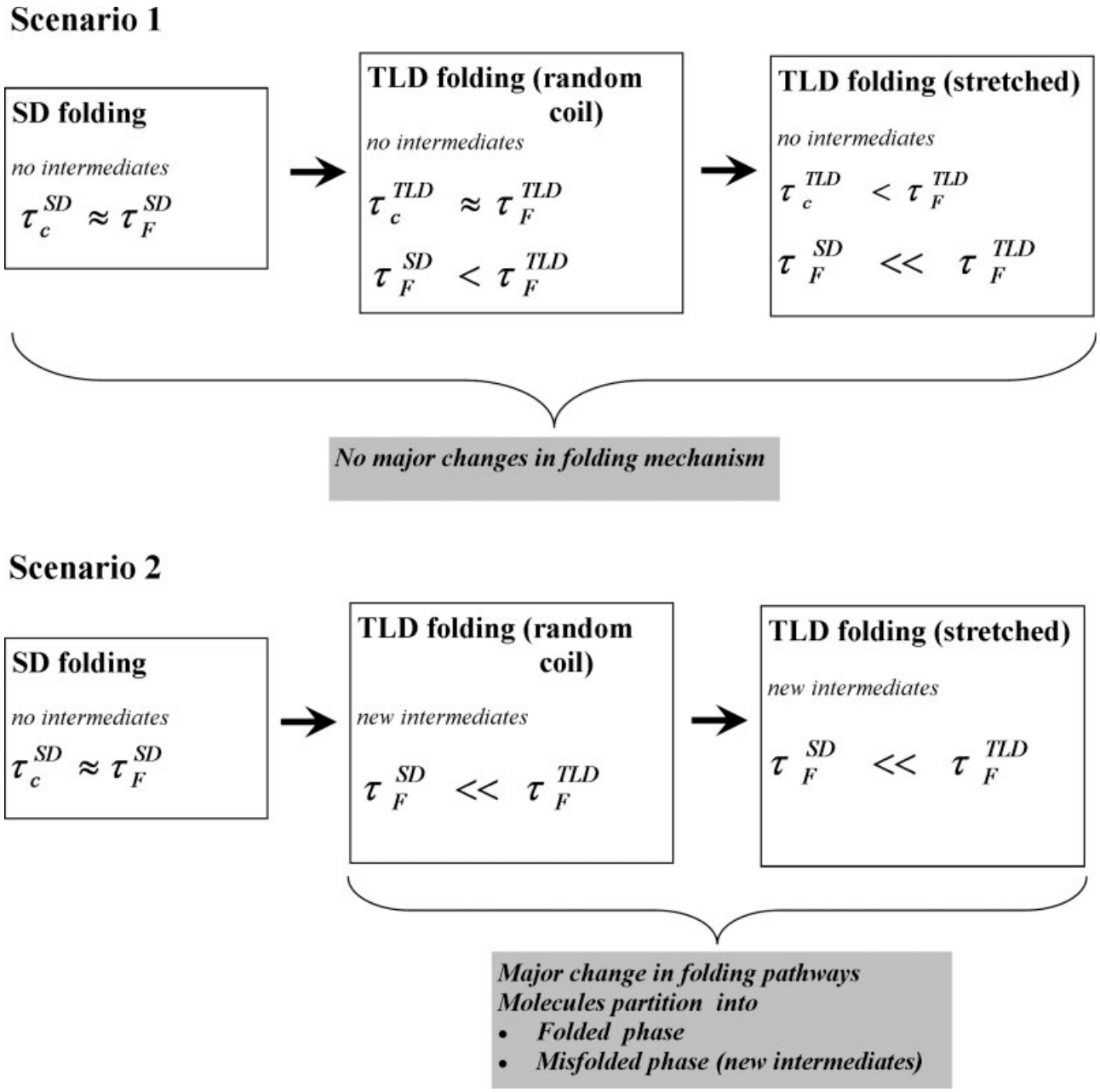

Fig. 7. The summary of tentative TLD folding scenarios for two-state folders. Random coil and stretched states are the different initial conditions for folding.

Both of these factors minimize the impact of tandem connectivity on S2 folding. In S2 domain, the terminal strand β4 has low stability and the β1–β4 interface is subject to fluctuations [Fig. 5(b)]. Furthermore, formation of stable native interactions associated with β4 is delayed compared to the growth of nucleation contacts elsewhere in the S2 fold (Fig. 6). As a result, β4 and its interface with β1 are poorly structured in the SD folding TSE. A highly ordered and rigid native core composed of the strands β1, β2, and β3 [Fig. 1(b)] provides a "template" for subsequent docking of β4 consistent with domain linkage.

The native state and TSE of S1 have the opposite characteristics. The terminal strands β1 and β4 represent the most stable and rigid part of S1 native core [Figs. 1(a) and 5(a)]. The interactions associated with β1 and β4, including their interface, are ordered in the TSE. Consequently, stable β1–β4 interface restricts the ability of S1 to fold consistently with tandem linkage. Our simulations suggest that efficient TLD folding

depends on relative stabilities of the interactions between terminals and elsewhere in the structure. Therefore, the factors identified above are likely to be related and determined by native energetics.

Limitations

It is important to discuss the scope of applicability of the factors facilitating TLD folding. First, these factors are deduced from the study of coarse grained models, which represent well the generic connectivity and topology of proteins. However, because of their simplicity these models cannot fully account for the diversity of protein interactions. Therefore, the S2-like features minimizing the impact of tandem linkage may not be always sufficient to avoid misfolding in tandems. These features may be overridden by sequence specific interactions. Second, current analysis of the impact of tandem linkage is limited to two-state folding domains. Third, results of our simulations and the proposed factors are applicable to the domains, in which terminal SSEs are in close proximity. Recent survey of small single-domain PDB proteins showed that approximately half of protein domains falls into this category.[31] Fourth, the impact of tandem connectivity is difficult to predict for the domains containing large number of SSE, N_{SSE}. To obtain a rough estimate of the number of domains, for which our results are likely to be applicable, we note that the distribution of protein domains with respect to N_{SSE} peaks at $N_{SSE} = 5$ or 6.[31] Assuming that the maximum N_{SSE}, to which the results of our study are applicable, is eight (as in Ig-like domains)[†], we estimate that approximately 40% of domains in the dataset considered by Krishna and Englander are sufficiently small. Therefore, our analysis is relevant for relatively large share of proteins.

Wild-type Domains

The natural questions, which arise from our simulations, are as follows: Are wild-type protein domains adopted to TLD folding? Are there specific domains, which are likely to misfold when incorporated into tandems? To answer these questions, we examined the native energetics and folding mechanisms of several proteins, for which detailed Φ-value data have been reported. Because wide sequence coverage Φ-value data are available for relatively few tandem-linked domains,[32–37] our objective is to demonstrate that SD folding mechanism of some tandem-linked wild-type domains is approximately S2-like. In particular, we selected two immunoglobulin (Ig)-like domains, I27 from titin and a fibronectin type III domain from tenascin, TNFn3. We also discuss the folding of two monomeric proteins Im9 and ACBP, which are not involved in wild-type tandems.

The energetics of wild-type proteins is evaluated by computing the interactions between SSE in the energy minimized native conformations using CHARMM22 force field. The fluctuations in their native structures at experimental conditions are probed using two approaches (Methods). First, we compute the root mean-squared displacement (RMSD) δR_i [Eq. (3)] for all residues i using all-atom explicit solvent molecular dynamics (MD). Second, from MD trajectories we obtain the standard deviations for the distances between SSE s_1 and s_2 $\delta R(s_1, s_2)$ [Eq. (4)]. A qualitative comparison between a wild-type domain and S1 (or S2) is based on the determination if the two factors facilitating TLD folding are applicable to a wild-type domain. Specifically, we consider

1. relative strengths of native SSE interactions;
2. flexibility of SSE represented by δR_i and $\delta R(s_1, s_2)$;
3. characteristics of TSE given by Φ-values.

The characteristics of S1 and S2 relevant for the comparison with wild-type domains are given in the Methods.

Immunoglobulin domain I27 from titin

The folding pathway of I27 has been investigated.[38] Detailed structural information about TSE has been recently reported using restrained MD simulations.[39] These studies showed that the folding nucleus consists of the core β-strands B, C, E, and F, which include the residues with the highest Φ-values [Fig. 8(a)]. MD sampling of TSE[39] further indicates that there are no direct interactions between A′ and G and the strand A appears to be disattached from the folded core. Our computations of RMSD values δR_i demonstrate that the most rigid segments of I27 largely coincide with the most structured TSE regions [Fig. 8(a)]. The average δR_i for the residues in the strands B, C, E, and F are 0.57, 1.28, 0.55, and 0.93 Å (the total average over B, C, E, and F is 0.84 Å). Furthermore, the most "rigid" residues, i.e. those with the minimal δR_i, are found within the strands B, E, and F. In contrast, the average RMSD for the residues in A, A′, and G are 1.07, 0.97, and 1.13 Å, respectively. The fluctuations in terminal strands are about twice larger than the fluctuations in nucleus strands B and E.

In the native state of I27, the lowest energy of nonbonded interactions (−186 kcal/mol) is between the strands B and E. The most rigid interfaces are between the strands B and E (δR(B, E) ≈ 0.09 Å) and between C and F (δR(C, F) ≈ 0.13 Å). For comparison, the distances between the terminal strands A, A′, and G experience larger fluctuations (δR(A′, G) ≈ 0.19 Å and δR(A, G) ≈ 0.21 Å). The unusual feature of I27 is that the N-terminal strand is split into two fragments, A and A′, which act as a clamp locking the two β-sheets. Consequently, in the native state the energy of A′-G interactions is low (−149 kcal/mol). Nevertheless, RMSD and the fluctuations in interstrand distances indicate high flexibility of terminal strands.

The analysis presented above suggests that there are qualitative similarities in the TSE structure and native energetics of I27 and S2. In both domains, the interface between terminal SSEs is excluded from the rigid native

[†]The limit $N_{SSE_{ESSE}} = 8$ is selected, because Ig-like domains seem to use both S2-like factors minimizing the impact of tandem linkage.

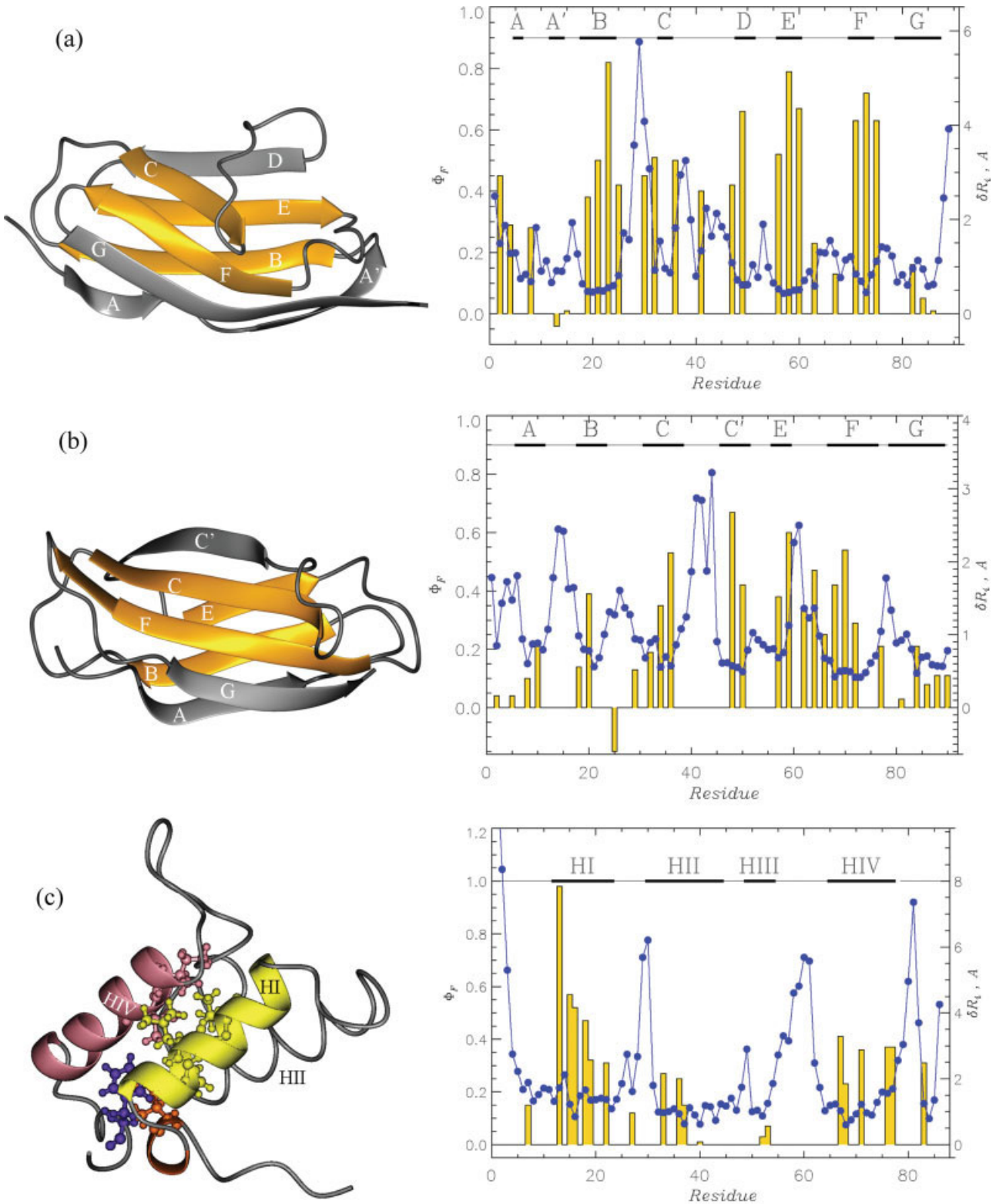

Fig. 8. Distributions of experimental Φ-values (bars) and computed RMSD δR_i [Eq. (3), circles] for I27[21] (a), TNFn3[32] (b), and Im9[40] (c). The value of δR_i measures the fluctuations in the position of residue *i*. Distributions of Φ-values, δR_i, and $\delta R(s1, s2)$ [Eq. (4)] are used to determine if a domain utilizes the factors, which minimize the impact of tandem connectivity on folding. Because in the TSE of I27 and TNFn3 the native core BCEF (in yellow) is structured and the terminal SSE are disordered, the folding of these domains is not significantly affected by tandem linkage. In Im9 the interface between terminal helices HI and HIV (in yellow and pink) is rigid in the native state and well structured in the TSE. Therefore, Im9 is likely to misfold when inserted in multidomain tandem. The short 3_{10}-helix is in orange and the residues with high Φ values are explicitly shown. The positions of SSEs are indicated in the top areas of the plots. The pictures are created using MolMol.[41] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

core (composed of the strands β1, β2, β3 in S2 or B, C, E, F in I27). The terminal β-strands are subject to large conformational fluctuations. In both I27 and S2, the native core is structured in the TSE, whereas the terminals (β4 in S2 and A, A′, G in I27) are largely disordered. Therefore, I27 appears to utilize the factors, which minimize the impact of tandem connectivity on folding.

Fibronectin type III domain from tenascin

The folding TSE of Ig-like domain TNFn3 has been studied experimentally[32] and using restrained MD simulations.[42] Both studies show that the core of TNFn3 made of the strands B, C, E, F is structured in TSE, while the terminal strands A and, possibly, G are not part of folding nucleus [Fig. 8(b)]. Similar to I27, TNFn3 is characterized by a BCEF-type folding nucleus, which excludes N- and C-terminal β-strands.[39]

The lowest interstrand energies in TNFn3 native structure are for F-G (−191 kcal/mol) and C-F (−133 kcal/mol) strand pairs, of which C-F is a part of the folding nucleus. Importantly, the A-G interactions are weak (−2.2 kcal/mol). Qualitatively, the distribution of RMSD δR_i [Fig. 8(b)] is similar to that of I27. The most rigid β-strands are the core strands B, C, E, and F as well as C′, for which the average RMSD ranges from 0.53 (F) to 0.86 Å (E). The longest span of residues with minimal δR_i is located in the strand F. The average RMSD for the terminal A is higher (1.06 Å). Consistent with previous MD simulations[39] few positions in G are rigid and experience relatively small fluctuations [Fig. 8(b)]. The analysis of fluctuations in interstrand distances shows that the most rigid parts of TNFn3 are the pairs of strands forming the nucleus C-F and B-E (δR(C, F) ≈ 0.12 Å and δR(B, E) ≈ 0.13 Å). In contrast, one of the largest fluctuations are observed in the distance between terminal A and G (δR(A, G) ≈ 0.30 Å). These data imply that the most disordered part of TNFn3 is the N-terminal strand A and the A-G interface is highly unstable. In contrast, β-strands forming the folding nucleus are rigid. Therefore, as I27, TNFn3 qualitatively resembles S2 and appears to incorporate the factors minimizing the impact of tandem connectivity.

TLD folding of Ig-like domains

On the basis of the analysis presented above, we conjecture that the TLD folding of Ig-like domains is similar to their SD folding. According to Scenario 1 in Figure 7, domain connectivity is expected to increase folding timescales, especially when stretched initial states are used. This scenario is qualitatively consistent with several AFM experiments that probed TLD folding with stretched initial conditions. For example, the TLD refolding rate for Ig-like 10Fn3 inserted in multidomain tandem is $k_F^{TLD} = 0.9\ s^{-1}$. In contrast, the rate of SD refolding is more than two orders of magnitude faster ($k_F^{SD} = 240\ s^{-1}$).[43] The decrease in k_F^{TLD} compared to k_F^{SD} has also been reported for I27.[44] For untethered I27 the refolding rate in water is $k_F^{SD} = 32\ s^{-1}$,[43] while k_F^{TLD} is more than 10 times slower ($1.3\ s^{-1}$).[44] Qualitatively similar kinetic consequences of tethering domain terminals by DNA molecular "handlers" were reported for RNase H.[13]

Scenario 1 for TLD folding initiated with random coil conformations predicts a moderate increase in folding timescales compared to τ_F^{SD}. These results agree well with the TLD and SD experiments on Ig[11] and TNFn3[12] tandem constructs, in which folding from random coils was monitored. For example, for most Ig domains in wild-type titin tandems, the slowdown in folding is about twofold compared to isolated single domains.[11] Qualitatively similar conclusions were drawn for TNFn3 domains.[12] More importantly, the experiments performed for isolated TNFn3 domains and those incorporated in the tandems revealed almost identical Φ-value distributions.[12] This implies that the TNFn3 folding mechanism remains largely intact despite tethering of domains' terminals. These experimental results support the Scenario 1 in Figure 7 and qualitative similarity between Ig-like domains and S2.

To illustrate that folding of single domain proteins may not be adapted to tandem connectivity, we consider bacterial immunity protein Im9 and acyl-coenzyme A-binding protein (ACBP). In contrast to Ig-like domains these proteins are not found in wild-type tandems.

Bacterial immunity protein Im9

The folding TSE of Im9, which folds via two-state mechanism, has been characterized by Φ-value analysis[40] and restrained MD.[45] These studies showed that the terminal helices HI and HIV are well structured in the TSE [Fig. 8(c)]. Importantly, the highest Φ-values are reported for the HI-HIV interface that implies native-like packing of hydrophobic interactions between two terminal SSE [Fig. 8(c)]. Experimental Φ-values[40] and accessible surface areas computed for individual residues[45] suggest that the middle helices HII and HIII are significantly less structured in the TSE than HI or HIV.

The analysis of native energetics shows that the HI-HIV interface has the lowest energy (−89 kcal/mol). The interactions between other helices are weaker, at least, by a factor of two, e.g., the next most stable interactions are between HII and HIII (−47 kcal/mol). Computation of δR_i indicates that the terminal helices HI and HIV are rigid in the native state [Fig. 8(c)]. Their average RMSD are 1.45 and 1.19 Å. The helices HII and HIII experience larger fluctuations and their average RMSD are 1.90[‡] and 1.65 Å, respectively. A 10-residue HIII-HIV loop is also highly flexible with the average RMSD of 3.82 Å. Besides HI-HIV interface, the C- and N-terminal interactions involve hydrophobic contacts between a short 3_{10} helix (residues 7–9) and Phe83 as well as a backbone hydrogen bond between Tyr10 and Lys84. As a result δR_i for $i = 8, 9$ are below 1.5 Å and one of lowest

[‡]To a large extend, the fluctuations in HII are due to the mobility of its N-termini.

δR_i(= 0.79 Å) corresponds to the position i = 84 [Fig. 8(c)]. Furthermore, the smallest fluctuations in interhelix distances are attributed to the terminal HI-HIV pair (δR(HI, HIV) = 0.19 Å). The fluctuations in other interhelix distances are significantly larger (e.g., δR(HII, HIII) = 0.37 Å). Therefore, N- and C-terminals constitute the most rigid segments in Im9 native fold, whereas the region between HII and HIV is flexible.

This analysis suggests qualitative similarities between Im9 and S1 domain. It appears that Im9 does not utilize the two factors, which minimize the impact of tandem connectivity. According to Scenario 2 in Figure 7, Im9 TLD folding is expected to be not only slower than SD folding, but be also qualitatively different due to formation of new misfolding intermediates. Better ordering of Im9 terminals in the TSE relative to other SSE may limit the ability of Im9 to adjust to tandem connectivity during folding. Because the native conformations and TSE of Im7 and Im9 are simular,[40,46,47] we propose that TLD folding of Im7 may also be susceptible to tandem linkage. TLD folding experiments can test this prediction.

Acyl-coenzyme A-binding protein (ACBP)

The site-directed mutagenesis probing the rate-limiting step in a two-state folding four-helix bundle ACBP has been reported.[48] The rate-limiting interactions in ACBP are found to be polarized and confined to the interface between the terminal helices A1 and A4. Specifically, single mutations of eight hydrophobic residues in the A1–A4 interface, each of which reduces the hydrophobic effect, were shown to slow down folding. In contrast, for the mutations at 18 positions spread over other ACBP regions including helices A2 and A3 the decrease in folding rate was not observed. These data suggest that the A1–A4 interactions are well ordered in the TSE, but other regions are poorly structured.

The computations of the native energetics indicate that the strongest attractive interactions are formed between A2 and A4 (−148 kcal/mol) and between A1 and A4 (−112 kcal/mol). The helix A3 interacts only with A2 and forms no contacts with A1 or A4. The distribution of RMSD values δR_i shows that the most rigid regions of ACBP are A1 (the average δR_i is 1.48 Å), A2 (0.94 Å), and A4 (1.14 Å). The helix A3 is characterized by significant fluctuations (1.88 Å). It follows from the computation of the fluctuations in interhelical distances that the A1–A4 interface is rigid (δR(A1, A4) = 0.17 Å). The fluctuations in other interhelix distances, particularly associated with A3, are larger (e.g., δR(A1, A3) = 0.40 Å). Because the interactions between terminal helices are well structured in the TSE and there is a strong A1–A4 coupling in the native state, ACBP qualitatively resembles S1 domain. As for Im9, we propose that ACBP is vulnerable to tandem linkage and may misfold once inserted into tandems.

Repeat proteins

Several recent studies investigated folding of TRP and ankyrin repeat proteins.[49–51] These proteins are constructed of multiple copies of double helix motif, which is repeated along a sequence. Because isolated individual repeats are unstable, the stability of repeat construct is drawn from extensive interactions between the repeats adjacent along the sequence. Because of the absence of long-range tertiary interactions, folding of repeat proteins is well described by 1D Ising model[49,51] and proceeds by adding repeats to growing folded phase. From this perspective, the repeat proteins are very different from the tandems of globular domains, which are stable and capable of folding as single isolated units.[12] Although the applicability of our findings to repeat proteins requires further studies, we still can consider as a "domain" a two-repeat unit, which represents a minimal stable construct.[49,50] The lack of interactions between the terminal helices in the two-repeat "domain" native state and, presumably, TSE is consistent with the Scenario 1 (Fig. 7). Ultrafast folding of repeat proteins and the absence of misfolding support the Scenario 1 predictions.[50]

CONCLUSIONS

Using coarse grained protein models, we examined the folding of tandem-linked domains and compared it with the folding of single isolated domains. Our results suggest that, in general, there can be two outcomes of linking domains into tandems (Fig. 7). The first scenario implicates rather modest impact on folding of tandem connectivity and, consequently, the mechanisms of folding of tandem-linked and single domains remain similar. According to the second scenario, tandem linkage leads to dramatic changes in folding mechanism. In particular, protein domains, which fold without intermediates as single domains, may misfold when inserted into tandems. The misfolded intermediates are created by topological constraints imposed by interdomain linkers. We found that tandem linkage always slows down folding.

By analyzing misfolding in model tandems, we suggest that the impact of tandem connectivity can be minimized by flexible terminal SSE. Specifically, two factors are proposed to facilitate TLD folding: (1) the interactions between terminal SSE are poorly ordered in the folding TSE, whereas nonterminal SSE are better structured, (2) the interactions between terminal SSE are weak in the native state. Our study suggests that Ig-like domains appear to utilize both of these factors. Consequently, tandem connectivity is likely to have a modest impact on Ig folding (Scenario 1 in Fig. 7). This conclusion of our study is in agreement with the experiments,[11,12] which showed that the main characteristics of SD folding for Ig-like domains are preserved in TLD folding. Therefore, bulk folding studies of isolated single Ig-like domains are likely to be applicable to TLD folding. Perhaps, it is not unexpected that we found Ig-like domains, which participate in wild-type multidomain constructs, to be adapted to tandem linkage. Our survey of the wild-type tandem-linked domains, for which Φ-value data are available,[21,32–37] suggests that S1-like design and, hence,

Scenario 2 are generally disfavored. Figure 8 also demonstrates that high Φ-values generally occur for the residues, which are most rigid in the native state. This observation suggests that native energetics largely determines the Φ-value distributions.

We are not aware of the experiments testing the validity of Scenario 2 in Figure 7. However, we propose several single domain proteins as possible candidates for this Scenario. Because these domains do not utilize the factors, minimizing the impact of tandem connectivity, their TLD folding could be qualitatively different from SD folding. Since terminal SSE are already "locked" in their TSE, errors in the position of interdomain linkers cannot be easily corrected. This might result in the formation of new category of folding intermediates, which do not occur in "normal" SD folding. Consequently, their TLD folding may partition into fast and slow kinetic phases, resembling kinetic partitioning in slow folding monomeric proteins.[28–30] The likely candidates for such scenario are single domain proteins Im9 and ACBP. Our findings also suggest that the main factor, which differentiates SD and TLD folding, is the topological effect imposed by interdomain linkers.

Because of the limitations of coarse grained models the proposed scenarios provide a simplified description of the consequences of tandem connectivity. In particular, their relevance to the proteins with large number of SSEs is unclear. More importantly, our model neglects attractive interdomain interactions, which are known to contribute to tandem stability[12,52] and dynamics.[53] Nevertheless, we believe that the proposed scenarios in Figure 7 capture the basic trends in the folding of tandem linked domains. The factors facilitating TLD folding can be used to rationalize and predict the effects of tandem connectivity. Future experiments and simulations using detailed models of wild-type tandems will further test the results reported here.

ACKNOWLEDGMENTS

The use of Altix SGI cluster for parallel computations at George Mason University is gratefully acknowledged. We thank Prof. Estela Blaisten-Barojas for valuable suggestions concerning this work.

REFERENCES

1. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
2. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biochem Sci 1999;24:77–83.
3. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
4. Labeit S, Kolmerer B. Titins, the giant proteins in charge of muscle ultrastructure and elasticity. Science 1995;270:293–296.
5. Geiger B, Bershadsky A, Pankov R, Yamada KM. Transmembrane crosstalk between the extracellular matrix and the cytoskeleton. Nat Rev Mol Cell Biol 2001;2:793–805.
6. Stossel TP, Condeelis J, Cooley L, Hartwig JH, Noegel A, Schleicher M, Shapiro SS. Filamins as integrators of cell mechanics and signalling. Nat Rev Mol Cell Biol 2001;2:138–145.
7. Ohashi T, Kiehart DP, Erickson HP. Dynamics and elasticity of the fibronectin matrix in living cell culture visualized by fibronectin-green fluorescent protein. Proc Natl Acad Sci USA 1999;96:2153–2158.
8. Baneyx G, Baugh L, Vogel V. Fibronectin extension and unfolding within cell matrix fibrils controlled by cytoskeletal tension. Proc Natl Acad Sci USA 2002;99:5139–5143.
9. Li H, Oberhouser A, Fowler S, Clarke J, Fernandez J. Atomic force microscopy reveals the mechanical design of a modular protein. Proc Natl Acad Sci USA 2000;97:6527–6531.
10. Oberhauser AF, Badilla-Fernandez C, Carrion-Vazquez M, Fernandez JM. The mechanical hierarchies of fibronectin observed with single-molecule AFM. J Mol Biol 2002;319:433–447.
11. Scott KA, Steward A, Fowler SB, Clarke J. Titin: a multidomain protein that behaves as the sum of its parts. J Mol Biol 2002;315:819–829.
12. Rounsevell RW, Stewart A, Clarke J. Biophysical investigations of engineered polyproteins: implications for force data. Biophys J 2005;88:2022–2029.
13. Cecconi C, Shank EA, Bustamante C, Marqusee S. Direct observation of the three-state folding of a single protein molecule. Science 2005;309:2057–2060.
14. Li MS, Hu C-K, Klimov DK, Thirumalai D. Multiple stepwise refolding of immunoglobulin domain I27 upon force quench depends on initial conditions. Proc Natl Acad Sci USA 2006; 103:93–98.
15. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding – A perspective from simple exact models. Protein Sci 1995;4:561–602.
16. Nguyen HD, Hall CK. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. Proc Natl Acad Sci USA 2004;101:16180–16185.
17. Klimov DK, Thirumalai D. Native topology determines force-induced unfolding pathways in globular proteins. Proc Natl Acad Sci USA 2000;97:7254–7259.
18. Veitshans T, Klimov DK, Thirumalai D. Protein folding kinetics: time scales, pathways, and energy landscapes in terms of sequence dependent properties. Fold Des 1997;2:1–22.
19. Camacho CJ, Thirumalai D. Kinetics and thermodynamics of folding in model proteins. Proc Natl Acad Sci USA 1993;90: 6369–6372.
20. Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. NAMD2: greater scalability for parallel molecular dynamics J Comput Phys 1999;151:283–312.
21. Fowler SB, Clarke J. Mapping the folding pathway of an immunoglobulin domain: structural detail from Φ value analysis and movement of the transition state. Structure 2001;9:355–366.
22. Klimov DK, Thirumalai D. Progressing from folding trajectories to transition state ensemble in proteins. Chem Phys 2004;307: 251–258.
23. Klimov DK, Thirumalai D. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. J Mol Biol 1998;282:471–492.
24. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich EI. On the transition coordinate for protein folding. J Chem Phys 1998;108:334–350.
25. Fersht AR. Characterizing transition states in protein folding. Curr Opin Struct Biol 1995;5:79–84.
26. Wolynes PG. Folding nucleus and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci USA 1997;94:6170–6175.
27. Shoemaker BA, Wang J, Wolynes PG. Structural correlations in protein folding funnels. Proc Natl Acad Sci USA 1997;94:777–782.
28. Thirumalai D. From minimal models to real proteins: time scales for protein folding. J Phys I 1995;5:1457–1467.
29. Kiefhaber T. Kinetic traps in lysozyme folding. Proc Natl Acad Sci USA 1995;92:9029–9033.
30. Matagne A, Radford S, Dobson C. Fast and slow tracks in lysozyme folding: insight into the role of domains in the folding process. J Mol Biol 1997;267:1068–1074.
31. Krishna MMG, Englander SW. The N-terminal to C-terminal motif in protein folding and function. Proc Natl Acad Sci USA 2005;102:1053–1058.
32. Hamill SJ, Stewart A, Clarke J. The folding of immunoglobulin-like greek key protein is defined by a common-core nucleus

andregions constrained by topology. J Mol Biol 2000;297:165–178.
33. Cota E, Steward A, Fowler SB, Clarke J. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. J Mol Biol 2001;305:1185–1194.
34. Sato S, Religa TL, Daggett V, Fersht AR. Testing protein-folding simulations by experiment: B domain of protein A. Proc Natl Acad Sci USA 2004;101:6952–6956.
35. Martinez JC, Serrano L. The folding transition states between SH3 domains is conformationally restricted and evolutionary conserved. Nat Struct Biol 1999;6:1010–1016.
36. McCallister E, Alm E, Baker D. Critical role of β-hairpin in protein G folding. Nat Struct Biol 2000;7:669–673.
37. Jager M, Nguyen H, Crane JC, Kelly JW, Gruebele M. The folding mechanism of a β-sheet: the WW domain. J Mol Biol 2001;311:373–393.
38. Wright CF, Lindorff-Larsen K, Randles LG, Clarke J. Parallel protein-unfolding pathways revealed and mapped. Nat Struct Biol 2003;10:658–662.
39. Geierhaas CD, Paci E, Vendruscolo M, Clarke J. Comparison of the transition states for folding of two Ig-like proteins from different superfamilies. J Mol Biol 2004;343:1111–1123.
40. Friel CT, Capaldi AP, Radford SE. Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. J Mol Biol 2003;326:293–305.
41. Koradi R, Billeter M, Wuthrich K. Molmol: A program for display and analysis of macromolecular structures. J Mol Graph 1996; 14:51–55.
42. Paci E, Clarke J, Steward A, Vendruscolo M, Karplus M. Self-consistent determination of the transition state for protein folding: application to a fibronectin type III domain. Proc Natl Acad Sci USA 2003;100:394–399.
43. Clarke J, Cota E, Fowler S, Hamill S. Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. Struct Fold Des 1999;7:1145–1153.
44. Carrion-Vazquez M, Oberhauser AF, Fowler SB, Marszalek PE, Broedel SE, Clarke J, Fernandez JM. Mechanical and chemical unfolding of a single protein: a comparison. Proc Natl Acad Sci USA 1999;96:3694–3699.
45. Paci E, Friel CT, Lindorff-Larsen K, Radford SE, Karplus M, Vendruscolo M. Comparison of the transition state ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with ϕ value restraints. Proteins Struct Funct Bioinform 2004;54:513–525.
46. Dennis CA, Videler H, Pauptit RA, Wallis R, James R, Moore GR, Kleanthous C. A structural comparison of the colicin immunity proteins Im7 and Im9 gives new insights into the molecular determinants of immunity-protein specificity. Biochem J 1998;333:183–191.
47. Capaldi AP, Kleanthous C, Radford SE. Im7 folding mechanism: misfolding on a path to the native state. Nat Struct Biol 2002; 9:209–216.
48. Kragelund BB, Osmark P, Neergaard TB, Schiodt J, Kristiansen K, Knudsen J, Poulsen FM. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. Nat Struct Biol 1999;6:594–601.
49. Mello CC, Barrick D. An experimentally determined protein folding energy landscape. Proc Natl Acad Sci USA 2004;101: 14102–14107.
50. Main ERG, Stott K, Jackson SE, Regan L. Local and long-range stability in tandemly arrayed tetratricopeptide repeats. Proc Natl Acad Sci USA 2005;102:5721–5726.
51. Kajander T, Cortajarena AL, Main ERG, Mochrie SGJ, Regan L. A new folding paradigm for repeat proteins. J Am Chem Soc 2005;127:10188–10190.
52. Litvinovich SV, Ingham KC. Interactions between type III domains in the 110 kDa cell-binding fragment of fibronectin. J Mol Biol 1995;248:611–626.
53. Sinha N, Kumar S, Nussinov R. Interdomain interactions in hinge-bending transitions. Structure 2001;9:1165–1181.