# Structural Comparison of the Two Alternative Transition States for Folding of TI I27

Christian D. Geierhaas,* Robert B. Best,* Emanuele Paci,† Michele Vendruscolo,‡ and Jane Clarke*
*Department of Chemistry, Medical Research Council Centre for Protein Engineering, University of Cambridge, Cambridge CB2 1EW, United Kingdom; †Institute of Molecular Biophysics, School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, United Kingdom; and ‡Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

ABSTRACT TI I27, a $\beta$-sandwich domain from the human muscle protein titin, has been shown to fold via two alternative pathways, which correspond to a change in the folding mechanism. Under physiological conditions, TI I27 folds by a classical nucleation-condensation mechanism (diffuse transition state), whereas at extreme conditions of temperature and denaturant it switches to having a polarized transition state. We have used experimental $\Phi$-values as restraints in ensemble-averaged molecular dynamics simulations to determine the ensembles of structures representing the two transition states. The comparison of these ensembles indicates that when native interactions are substantially weakened, a protein may still be able to fold if it can access an alternative transition state characterized by a much larger entropic contribution. Analysis of the probability distribution of $\Phi$-values derived from ensemble averaged simulations, enables us to identify residues that form contacts in some members of the ensemble but not in others illustrating that many interactions present in transition states are not strictly required for the successful completion of the folding process.

## INTRODUCTION

Considerable experimental evidence suggests that most of the proteins that fold by a two-state mechanism reach their native state via a single transition state (1–3). These proteins populate only two dominant free-energy minima and also often allow only one pathway to interconvert between these states. Such an evolutionary design may be a consequence of the necessity to minimize the tendency to misfold (4–8). It appears to be remarkably effective; indeed only a few proteins are known to fold by multiple folding pathways (9), the majority of those result from *cis*/*trans* prolyl isomerization (10,11) or from the presence of intermediates that might be en route to the native state or a nonproductive off-pathway trap (11–15). The existence of multiple unfolding pathways has also been proposed on the basis of theoretical considerations, e.g., (4,16–18). One of the most unambiguous experimental examples of a protein having two alternative folding pathways is the case of the immunoglobulin domain TI I27 (7,8).

In the case of two-state proteins, important insights about the determinants of the folding process have been provided by the evidence that the folding nucleus can be changed by mutations (e.g., protein G and protein L) (19) and by circular permutation (e.g., S6) (6). It is therefore of great interest to analyze in detail the case in which changes in the thermodynamic conditions, not simply in the sequence, can promote folding along alternative pathways.

It has recently been suggested that folding mechanisms fall on a continuum between strictly hierarchical, with early formation of secondary structure, to purely nucleation-condensation (concomitant formation of secondary and tertiary structure) and that the folding mechanism for any given protein will depend on the secondary structure propensity of the molecule (20). Where secondary structure propensity is high, a hierarchical mechanism (such as diffusion-collision) is likely, as has been observed in the engrailed homeodomain and protein G (20,21). By contrast, where secondary structure propensity is low, a pure nucleation-condensation mechanism will be observed, as shown classically for CI2 (22,23). The transition states observed in nucleation condensation mechanisms are diffuse, with a characteristic pattern of high and low fractional $\Phi$-values distributed throughout the molecule.

Polarized transition states have been observed in a number of proteins, probably representing a hybrid between nucleation-condensation and diffusion-collision folding mechanisms (20,24,25). High $\Phi$-values are not evenly distributed over the molecule but usually clustered at one position, with the rest of the structure exhibiting mainly low $\Phi$-values (hence they may also be called ''localized'' transition states). A polarized TSE is observed in SH3 domains,

*Submitted November 3, 2005, and accepted for publication February 14, 2006.*

Address reprint requests to Jane Clarke, Tel.: 44-1223-336426; Fax: 44-1223-336362; E-mail: jc162@cam.ac.uk; or Michele Vendruscolo, Tel.: 44-1223-763848; Fax: 44-1223-336362, E-mail: mv245@cam.ac.uk.

Robert B. Best's current address is Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda, MD 20892-0520.

*Abbreviations used*: TS, transition state; TSE, transition state ensemble; $TS_L$, transition state of TI I27 dominant at low concentration of denaturant; $TS_H$, transition state of TI I27 dominant at high concentration of denaturant; $TSE_L$, transition state ensemble of TI I27 dominant at low concentration of denaturant; $TSE_H$, transition state ensemble of TI I27 dominant at high concentration of denaturant; pathway L, folding pathway of TI I27 dominant at low concentration of denaturant; pathway H, folding pathway of TI I27 dominant at high concentration of denaturant; *SASA*, solvent accessible surface area; *Rg*, radius of gyration; RMSD, root mean-square deviation; dRMS, distance-based root-mean-square deviation; EEF1, effective energy function; and Ig, immunoglobulin.



0006-3495/06/07/263/13 $2.00 doi: 10.1529/biophysj.105.077057

in which fully established contacts are mostly observed in turns that have to be formed to bring the remaining chain together (26–30). A polarized transition state is not necessarily unstructured: Garcia-Mira et al. recently examined the transition state for folding of CspB using Φ-value analysis (25). CspB folds via a strongly polarized transition state with a $\beta_T$ of 0.9, thus most regions of the protein are folded in the TS—but most native interactions are not yet fully established (25). The most polarized Φ-value distribution, and thus transition state, has been observed in the circular permutant *P*13-14 of S6 (6). This result is particularly interesting because wild-type S6 usually folds via a nucleation-condensation mechanism (31,32). In the circular permutant, the folding nucleus is polarized toward the artificially linked C- and N-termini.

In this work we investigate whether the different patterns of Φ-values observed for TI I27 under different experimental regimes indicate a switch in folding mechanism, brought about by changes in conditions, rather than a change in sequence or chain connectivity. We analyze the two transition states for folding of TI I27 that have been characterized individually by Φ-value analysis (7). We use the experimental Φ-values of either transition state as restraints in molecular dynamics simulations. The simultaneous enforcement of a large number of restraints derived from experimental Φ-values in molecular dynamics simulations not only results in an ensemble of conformations that represents the transition state, but also avoids possible problems in the interpretation of individual Φ-values (33,34). The fact that a single polypeptide chain is required to satisfy all the restraints simultaneously considerably reduces the number of alternative explanations that should be considered when one considers individual Φ-values on their own. The remarkable success of this type of restrained simulation is a consequence of the nucleation mechanism, which requires the formation of a specific set of interactions in the transition state. Therefore the topology of the transition state is determined when just a small number of Φ-values are specified (3,35–38).

The comparison of the two corresponding transition state ensembles of TI I27 indicates that at physiological conditions (denoted as pathway L), this protein folds by a classical nucleation-condensation mechanism, characterized by a tightly packed folding nucleus with the remaining structure in a nativelike topology, whereas at high concentrations of denaturant (denoted as pathway H) it follows a different mechanism, characterized by a polarized transition state.

## METHODS

### Protein system

The 89-residue protein TI I27, the 27th immunoglobulin domain from human cardiac titin, exhibits a $\beta$-sandwich structure with two $\beta$-sheets. The A-B-E-D sheet is the N-terminal sheet and the C-F-G-A′ sheet is the C-terminal sheet. As starting structure for all simulations we used the NMR solution structure of TI I27 (67), 1TIT in the Protein Data Bank (see Fig. 1).

The Φ-values were measured by Wright et al. using unfolding kinetics and extrapolated to 0 M GdmCl (7). For the pathway dominant at moderate conditions, i.e., via $TSE_L$, the same set of 22 Φ-values as used in our previous study were taken as input in simulations. The Φ-value of I23A was measured to be 1.22(±0.03), but in the simulations this value was set to 1.0 because a Φ-value >1.0 cannot be interpreted in our model using fractions of native contacts. Similarly, G32A was neglected because Φ-values for Gly are undefined in our simulations (38). In the pathway dominant at extremes of denaturant the negative Φ-values of V4A, L8A, H56A, and A75 were set to zero because the negative Φ-values cannot be interpreted in the model of calculated fractions of native contacts. We expect these residues to be nonnative in the transition state. The Φ-values for I23A, G32A, A82G, and V86A were neglected because they are either nonclassical Φ-values or associated with large errors and I2A was also set to zero because we expect it to be zero in $TSE_H$. Finally a set of 19 experimental Φ-values were used as input in simulations for the pathway dominant at extremes of denaturant, i.e., via $TSE_H$.

For TNfn3, residues 803–891 in the original Protein Data Bank entry 1TEN were renumbered 1–89 with L803 as residue 1. It is the same numbering as used in our previous simulations (35,36). This numbering differs from that used by Hamill et al. (52); they numbered residues 1–90 with the incompletely resolved R802 as residue 1.

### Equilibration of the native state

We used the same protocol for the equilibration of the native state described in Paci et al. (36). Simulations were performed with the program CHARMM (68) and with an all-atom energy function (EEF1) that implicitly includes the effects of the solvent (69).

Starting from the experimental structure (1TIT), we first performed a short steepest descent minimization to remove possible steric clashes. We then heated the protein to 300 K and equilibrated it for 0.5 ns. During equilibration the protein remained relatively close to the initial conformation. The resulting structure is the actual starting structure for the simulations to determine the TSE. The integration step was 2 fs for all of the simulations performed. Temperature was kept constant using the Nose-Hoover thermostat.

### Molecular dynamics simulations restrained with Φ-values

A pseudoenergy term was imposed on the MD simulations to guide the sampling toward transition state structures. In the simulations, the Φ-value of residue *i* in a particular configuration was defined as the fraction of native contacts made by that residue, i.e., $\Phi_i^{sim} = N_i/N_i^{nat}$. When more than one replica was used, the Φ-value was defined to be the average Φ-value over all replicas. To drive the simulations toward satisfying the experimental data, the biased molecular dynamics method of Paci and Karplus (70) was adapted for the replica method as follows: a reaction coordinate $\rho$ was defined as the mean-squared difference between the $N^e$ experimental Φ-values, $\Phi_i^{exp}$, and the simulated Φ-values, $\Phi_i^{cal}$ (Eq. 1):

$$\rho = \frac{1}{N^e}\sum_i (\Phi_i^{cal} - \Phi_i^{exp})^2. \quad (1)$$

To avoid forcing the system toward the desired low values of $\rho$, an energy was added to the Hamiltonian only when $\rho$ exceeded the lowest value, $\rho_0$, attained up to that point in the simulation, as defined by Eq. 2:

$$E = \begin{cases} \frac{\alpha M}{2}(\rho - \rho_0)^2 & \rho > \rho_0 \\ 0 & \rho \le \rho_0 \end{cases}. \quad (2)$$

In this expression, the constant $\alpha$ controls the weight of the bias relative to the force field. The factor *M*, corresponding to the number of replicas,

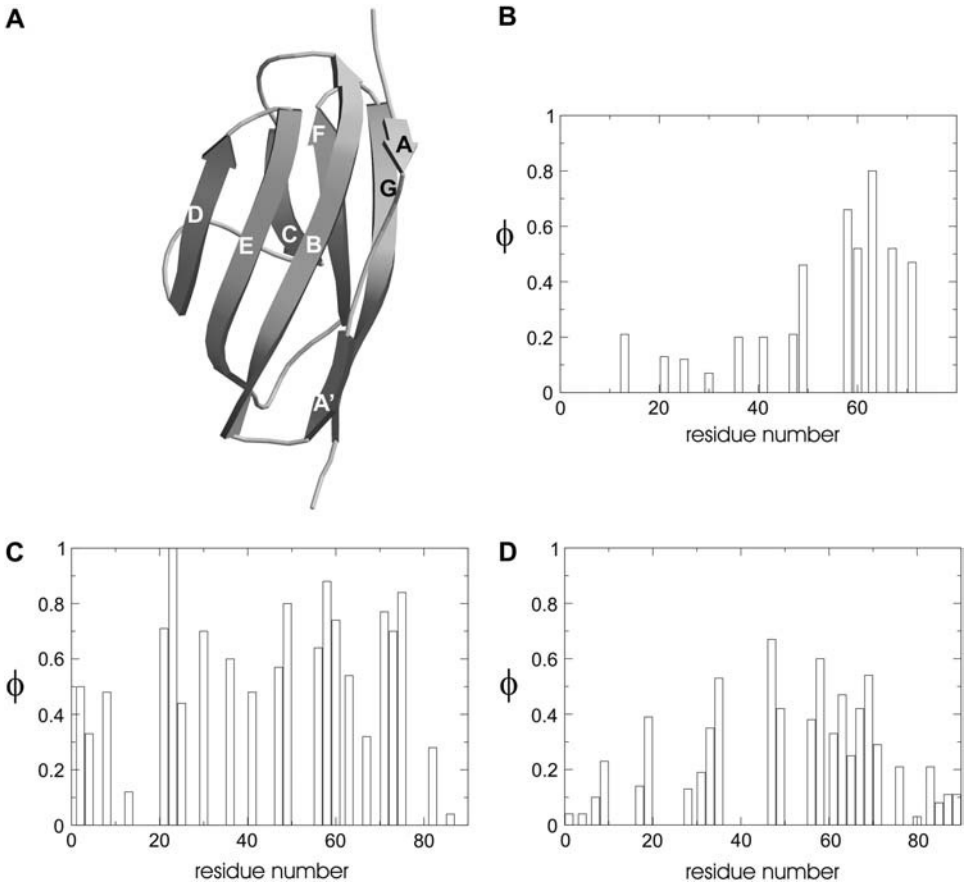


FIGURE 1 Structure of TI I27 and experimental Φ-values. (*A*) TI I27 has an Ig-like fold (67). The native structure has one $\beta$-sheet containing the A, B, D, and E strands and a second $\beta$-sheet with the C, F, G, and A′ strands. This picture was generated using the programs RASMOL (74), MOLSCRIPT (75), and RENDER (76). Experimental Φ-values of $TSE_H$ (*B*) (7), (*C*) $TSE_L$ (7,41), and (*D*) TNfn3 (52).

ensures that the restraints will have the same weight relative to the force-field energy, regardless of the number of replicas. Information was shared between the replicas using MPI routines, as described in more detail in Best et al. (71). Transition state ensembles consisting of one, two, or four replicas were generated by this method (72).

## Selection of the sampling temperature

A harmonic potential is applied to keep the reaction coordinate $\rho$ close to zero, so that the $\Phi^{exp}$ values are restrained around the $\Phi^{cal}$ values and the TSE is sampled at different temperatures. The prefactor $\alpha$ of the harmonic term was chosen so that $\rho(t)$ is kept at $<0.001$. Then the temperature is increased to enhance sampling efficiency and to favor the presence of higher energy, nonnative structures.

We have used the $\beta_T$-value to estimate *SASA* of the TSE of both $TS_L$ and $TS_H$/TNfn3 (C. D. Geierhaas,. A. A. Nickson, K. Lindorff-Larsen, J. Clarke, and M. Vendruscolo, unpublished data). The results are 7000 ± 800 and 5500 ± 500 Å$^2$ for $TS_H$/TNfn3 and $TS_L$, respectively. In the multireplica simulations for $TS_L$, the structures at 360 K showed the best agreement with the expected value of 5500 ± 500 Å$^2$. For $TS_H$ and TNfn3, the structures sampled at 500 K were in the range of the predicted value (7000 ± 800 Å$^2$). In the single replica case structures were sampled at 300, 360, 430, 500, 640, and 780 K and accepted if their solvent-accessible surface area was within the predicted region.

## Calculation of the interaction maps

The all-atom energy function (EEF1) that implicitly includes the effects of the solvent (69) can be decomposed into the sum of pairwise interactions between residues (69,73). Thus, the effective energy, which includes the solvent contribution, can be written as a sum over pairs of residues as:

$$E_{EEF1} = \sum_i \sum_{j \geq i} E_{ij}. \quad (3)$$

The energy map is a graphical representation of interaction matrices where the element $i$, $j$ is the EEF1 interaction energy between residues $i$ and $j$, averaged in the TSE and in the equilibrated native state. Only noncovalent interactions between amino acids that are at least two residues apart are considered. The B-score is calculated as defined by Vendruscolo et al. (39).

# RESULTS

## The transition state ensemble of pathway H ($TSE_H$)

The general properties of the ensembles of structures representing the transition state ensemble for the pathway dominant at high concentration of denaturant, $TSE_H$, are reported in Table 1. The four-replica ensemble is slightly more heterogeneous in terms of RMSD than the one- and two-replica ensembles, but they share other overall features, such as average solvent accessible surface area (*SASA*) and radius of gyration (*Rg*). The data that are discussed in the following sections were gathered from four-replica simulations, but similar results are obtained from the one- and two-replica simulations.

**TABLE 1 General properties of the transition state ensembles**

| Protein | $N$* | $N\langle\Phi^{exp}\rangle$† | RMSD (Å)‡ | $Rg$ (Å)§ | $\Delta Rg$§ | $S$ (Å$^2$)¶ | $\Delta S$¶ | $\langle\Phi^{cal}\rangle$‖ | $\langle\Phi^{exp}\rangle$** |
|---|---|---|---|---|---|---|---|---|---|
| 1TIT | 4 | 19 | 11.3 (2.2) | 14.5 | +12% | 7200 | +41% | 0.18 | 0.24†† |
| $TS_H$ | | | | (0.8) | | (500) | | | |
| 1TIT | 2 | 19 | 10.0 (1.3) | 14.5 | +12% | 7200 | +41% | 0.17 | 0.24†† |
| $TS_H$ | | | | (0.7) | | (400) | | | |
| 1TIT | 1 | 19 | 9.1 (1.5) | 14.1 | +9% | 7100 | +39% | 0.19 | 0.24†† |
| $TS_H$ | | | | (0.5) | | (300) | | | |
| 1TIT | 4 | 4‡‡ | 10.0 (3.5) | 14.4 | +11% | 7100 | +39% | 0.21 | 0.24†† |
| $TS_H$ | | | | (1.1) | | (700) | | | |
| 1TIT | 4 | 22 | 4.2 (0.7) | 12.8 | −1% | 5500 | +8% | 0.48 | 0.57†† |
| $TS_L$ | | | | (0.2) | | (200) | | | |
| 1TIT | 1 | 22 | 4.4 (0.7) | 12.9 | −0.5% | 5550 | +9% | 0.49 | 0.57†† |
| $TS_L$ | | | | (0.2) | | (250) | | | |
| 1TEN | 4 | 26 | 9.9 (3.2) | 14.9 | +14% | 7100 | +36% | 0.20 | 0.28§§ |
| | | | | (1.4) | | (600) | | | |

The number reported in parentheses corresponds to ±1 SD from the mean.
*Number of replicas.
†Number of experimental Φ-values used as restraints in the simulation.
‡RMSD is the mean Cα root mean-square distance from the native conformation.
§$Rg$ is the mean radius of gyration and $\Delta Rg$ is the difference in $Rg$ relative to the native state.
¶$S$ is the mean solvent-accessible surface area and $\Delta S$ is the difference in $S$ relative to the native state.
‖$\langle\Phi^{cal}\rangle$ is the average calculated Φ-value computed from all the residues that have nonzero number of side-chain native contacts.
**$\langle\Phi^{exp}\rangle$ is the average experimental Φ-value.
††Φ-values taken from Wright et al. (7).
‡‡Key residue simulation using only the experimental Φ-values from residues L58, L60, C63, M67.
§§Experimental Φ-values taken from Hamill et al. (52).

A total of 4000 structures were generated using 19 experimental Φ-values as restraints (7). Simulations were performed at a pseudotemperature of 500 K, because at this temperature the structures sampled have a β-Tanford value ($\beta_T$) close to the experimental value of 0.7 (see Methods). The $TSE_H$ is highly heterogenous, with an average Cα-RMSD of 11.3 ± 2.2 Å (see Table 1). The average *SASA* of the $TSE_H$ is 7200 ± 500 Å$^2$, an increase of 41% compared to the native state. In parallel to this increase in *SASA*, the $Rg$ increases to 14.5 ± 0.8 Å, 12% larger than the native state. The average over all residues (measured experimentally or not) of the calculated Φ-values, $\langle\Phi^{cal}\rangle$, is ~0.18, which is close to the experimental average $\langle\Phi^{exp}\rangle = 0.24$. A correlation coefficient of 0.99 between the experimental and the calculated Φ-values indicated that the procedure used to determine the transition state ensemble was self-consistent. Fig. 2 *A* shows the Φ-values for individual residues. The red curve is the average calculated value, whereas the blue dashed boundaries represent mean ±1 SD, respectively. The area between the dashed lines in the plot, therefore, characterizes the level of confidence with which Φ-values can be determined in the simulations. As expected from the low values of $\langle\Phi^{cal}\rangle$ and $\beta_T$, the $TSE_H$ is rather heterogeneous and contains several nonnative interactions, except in the region of residues L58–V71 (E strand, E-F loop, and F strand) and residue I49 (D strand), which have higher experimental Φ-values (>0.46) than the other residues. The residues in the G strand are also predicted to be rather structured, although reliable experimental Φ-values in this strand could not be measured. Thus we do not have sufficient confidence in these results for the G strand. (Note that in the absence of Φ-values to constrain the system away from the native state the force field may tend to maintain native contacts). The standard deviation of the calculated Φ-values is large in most regions of the protein as a consequence of the heterogeneous nature of $TSE_H$.

The average Cα-RMSD between pairs of structures is 12.5 Å indicating a substantial structural heterogeneity in the transition state ensemble. To analyze the structures in more detail, they were clustered together using a 3 Å cutoff (38). We obtained 430 groups, of which only nine have more than 20 members, as expected for a heterogeneous ensemble of structures. The cluster centers have very different properties, in particular $\Delta SASA$ ranges from +22% to +74% and $\Delta Rg$ ranges from 0% to +39%. Yet, all these structures fulfill the restraints of the experimental Φ-values. Representative cluster centers are shown in Fig. 3.

## Secondary structure and hydrogen bond patterns

To analyze the secondary structure content of the $TSE_H$ in detail we characterized the hydrogen bond patterns and the secondary structure content of the nine highest populated cluster centers—each of them with more than 20 members. This analysis (Fig. 4) suggests that $TSE_H$ consists of a partially formed B-E-D β-sheet and a partially formed C-F(-G) β-sheet. In the B-E-D sheet the contacts between strands E and D are more fully formed with an average of three native H-bonds between them. The B strand is attached more loosely (an average of one native H-bond). A number of

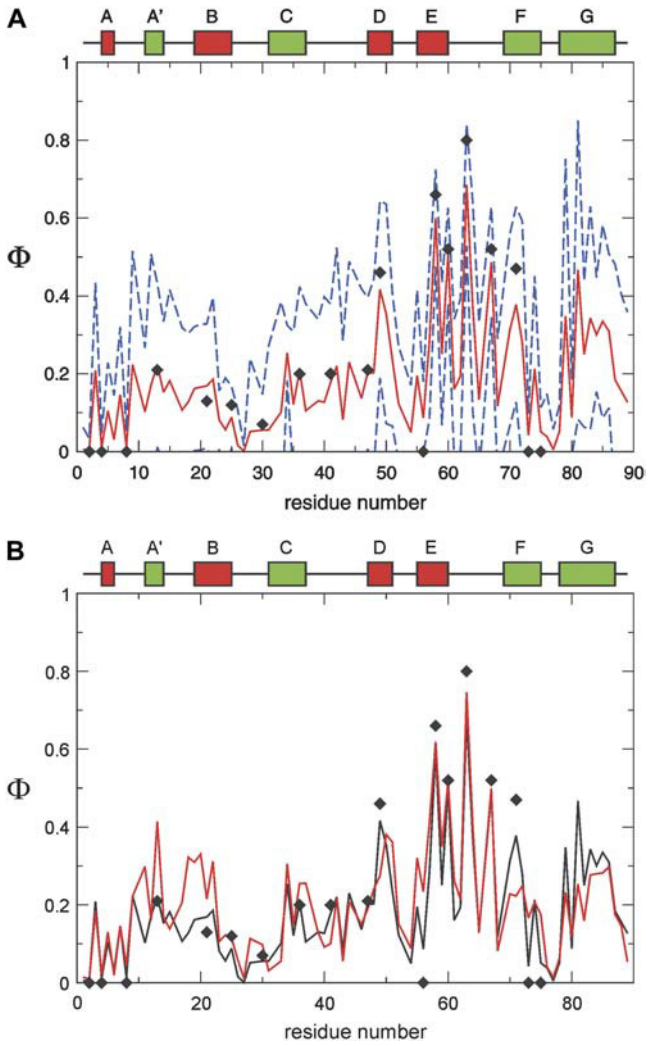


FIGURE 2 Experimental and calculated Φ-values for the $TSE_H$ of TI I27. (*A*) $TSE_H$ determined using the initial set of 19 $\Phi^{exp}$. The diamonds show the experimental Φ-values. The red line is the average calculated Φ-value, and the blue dashed lines represent the average ±1 SD within the $TSE_H$. The size of the white area between the dashed lines represents the confidence with which the Φ-values are predicted at any position where no experimental Φ-values are available. Most of the regions of the protein are highly heterogeneous. (*B*) Comparison of Φ-values calculated using either all 19 experimental Φ-values (*black curve*), or subset L58, L60, C63, and M67 (*red curve*).

nonnative H-bonds between the strands are also observed between stands B and E. The $\beta$-sheets are held together by the intersheet loops, especially the E-F loop (see also ''The structure of the loop regions'' in the Discussion section).

## Network of interactions in the transition state

To analyze the network of interactions in detail we calculated the pairwise interaction energies between all possible residue pairs for both transition states and for the native state. These energy maps are shown in Fig. 5, where the native states are represented above the main diagonal and the transition states below. Although the $TSE_H$ is rather unstructured, the overall topology of the network of interactions is still nativelike. Most of the nonnative interactions are weak compared to the interactions that establish the native topology of the Ig-like fold. From an analysis of hydrogen bonding patterns we proposed that in the B-E-D sheet contacts are more fully formed between strands E and D than between strands B and E. This result is also illustrated by the interaction map. Interactions between the E and D strands are strong, whereas interactions between the B and E strands are formed but significantly weaker. Weak C-F interactions are evident from the interaction maps. Again we ignore the G-strand contacts because we have no confidence that the simulations are correct. It is interesting to note that in the $TSE_H$ most of the short-range interactions are still strong, but the long-range interactions tend to be considerably weakened. Most of the A-B and A′-B interactions are lost, confirming the unstructured nature of A and A′.

## Key residue simulations

To determine if at extreme concentrations of denaturant the protein folds via a mechanism that involves key residues, we

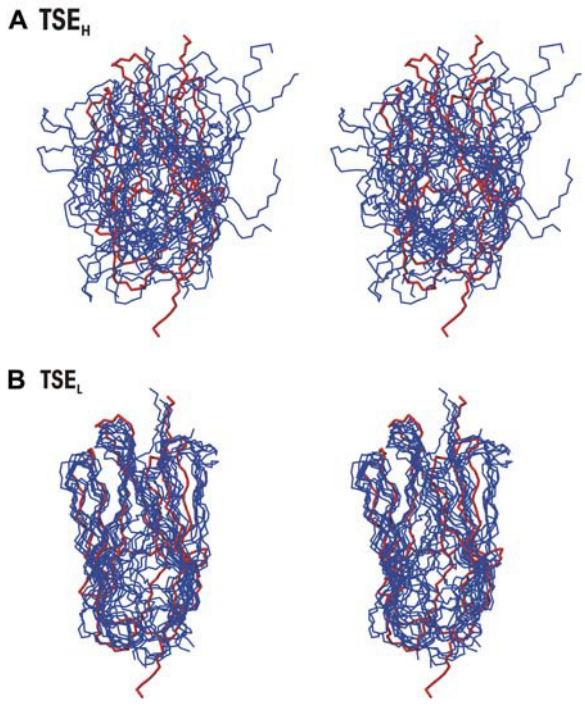


FIGURE 3 Representative structures of the transition state ensembles of $TSE_H$ and $TSE_L$. The structures that make up the $TSE_H$ were clustered using a RMSD of 3 Å as a cutoff. (*A*) Stereographic view of four representative structures (cluster centers) of the $TSE_H$ (*blue*, *thin lines*) and the native state (*red*, *thick line*). The structures are significantly unfolded and less structured than the native state. (*B*) Stereographic view of the eight highest populated representative structures (cluster centers) of the $TSE_L$ (*blue*, *thin lines*) and the native state (*red*, *thick line*). The structures are very nativelike compared to those of $TSE_H$. The figures were drawn with the program MOLMOL (77).

investigated whether a set of key residues according to Vendruscolo et al. (3) could be defined. Predictions made by using the network betweenness, the so-called B-score (39), indicated that residues F21, W34, L58, L60, M67, V71, F73, and L84 are the most central residues in the network of interactions in the $TSE_H$. But in key residue simulations (i.e., using only these residues as restraints) the Φ-value profile (the whole set of 19 experimental Φ-values) could not be reproduced accurately (data not shown). In contrast the set consisting of residues L58, L60, C63, and M67 can reproduce the Φ-value profile with high confidence. The coefficient of correlation to the full set of experimental Φ-values (excluding the key residues used as restraints) is 0.7 (and the coefficient of correlation including the key residues is 0.9). The agreement with the full set of simulated Φ-values is also good (correlation coefficients of 0.8 and 0.9 without and with the key residues, respectively). Major differences between $\Phi^{cal}$ from the key residue simulation and $\Phi^{cal}$ calculated from a simulation using the full set of 19 experimental Φ-values appear in the B strand (see Fig. 2 *B*). Additional properties such as *SASA*, *Rg*, and RMSD can be determined with confidence (see Table 1). Two other simulations were performed using different sets of four residues each excluding one of the key residues identified above, but including another residue with a high Φ-value (58,60, 63 plus 49, and 60,63,67 plus 71). In these cases the correlation with the experimental Φ-values was low (0.3 and 0.5, respectively).

### Ensemble averaged molecular dynamics simulations

The use of ensemble averaged molecular dynamics simulations allows comparison of the distribution of probabilities of Φ-values for each residue. The results are shown in Fig. 6 *A* for both $TSE_L$ and $TSE_H$.

Several residues have a bimodal distribution of Φ-values in $TSE_H$. Of particular interest are the cases of residues I49 and V71, both of which could be considered key according to their B-score (39). A simulation with the key residues L58, L60, C63, and M67 plus I49 and V71 produced similar results to the simulation that included L58, L60, C63, and M67 alone, suggesting that I49 and V71 do not belong to the key residue set. Both I49 and V71 have a unimodal distribution of Φ-values in the single-replica system, with a maximum at ~0.4. In the multireplica simulations, however, they exhibit a bimodal distribution of Φ-values, with two maxima at 0.3 and 0.7. This finding supports the idea that I49 and V71 are not key residues, and they can either participate in the folding process or not; hence they are not crucial for the successful establishment of the transition state (32,33,40). To demonstrate this behavior of I49 and V71 in more detail we computed the Φ-values of I49 and V71 for each structure in the TSE (2 replica system, Fig. 6 *B*). Most of the $\Phi_{I49}/\Phi_{V71}$ pairs either have low $\Phi_{I49}$ and high $\Phi_{V71}$ or vice versa. The structures that correspond to the data in the ellipsoids represent 75% of the whole ensemble. As expected, $\Phi_{I49}$ and $\Phi_{V71}$ are anticorrelated with a correlation coefficient of −0.75. This result was not observed using single replica simulations and demonstrates the advantages of ensemble averaged molecular dynamics simulations. Importantly, the key residues that are important in the formation of $TSE_H$ show a unimodal distribution of Φ-values even in the four-replica simulations.

In $TSE_L$ several of the Φ-values of the residues in the region of strands A and A′ have bimodal distributions, at least in the multireplica simulations. This result is in agreement with the finding that the A and A′ strands are much less structured than the B, C, D, E, and F strands in $TSE_L$ (35,41). They were found to be in a variety of conformations in $TSE_L$ and participate in several nonnative interactions. Most of the bimodal distributions of Φ-values for these residues have maxima at low values (Φ close to 0) and at high values ($\Phi \geq 0.6$). Thus they are either unstructured in $TSE_L$ or are already in their nativelike conformation. This result was observed for nine of the 15 residues in strands A and A′ in the multireplica case but only for three in the single-replica simulations. Another region with many bimodal distributions of Φ-values is the C-D loop (residues 37–46) with five bimodal distributions. Also in these cases these residues are either unstructured with a maximum close to $\Phi = 0$ or they are already significantly structured with $\Phi \geq 0.6$. Most of these residues exhibit native contacts with the large W34 in the hydrophobic nucleus of $TSE_L$. Thus, most residues in the C-D loop are only marginally involved in the folding nucleus and participate to a different extent in each folding event (32,33,40). This situation is comparable to the one observed in the similar simulations of the first transition state for folding of barnase in which an α-helix is formed or not (42). Importantly, residues 41 (C-D loop) and 76 (F-G loop) are very interesting, because their Φ-values are highly anticorrelated, with a correlation coefficient of −0.80. Thus, either 41 exhibits a high and 76 a low Φ-value or vice versa. Both residues pack into the core and are on opposite sites of it, hence either the C-D loop or the F-G loop pack onto the core to facilitate folding. Interestingly, most of the bimodal distributions of Φ-value probabilities in the C-D loop were also sampled by the single-replica system.

## DISCUSSION

### Comparison of the two TSEs of the parallel pathways

The behavior of TI I27 is very unusual, as it has been shown experimentally to fold via two different pathways and they have both been characterized by Φ-value analysis (7). In the following section we will investigate these remarkable differences in detail and will show that TI I27 can fold both via a polarized transition state and via a more common nucleation-condensation mechanism.

### *Overall properties*

The folding transition state of pathway L is very nativelike compared to the transition state for pathway H, as indicated by the comparison of the average experimental Φ-values (0.24 and 0.57 for $TS_H$ and $TS_L$, respectively) and the $\beta_T$ (0.74 and 0.95 for $TS_H$ and $TS_L$, respectively) (7,41). Crucially, the $TS_H$ is not just an expanded version of $TS_L$, as the patterns of Φ-values are very different.

The general properties of the $TSE_H$ and $TSE_L$ are reported in Table 1. To simplify the comparison between the ensembles we also determined $TSE_L$ using ensemble-averaged simulations. Although in $TSE_L$ the protein has a C$\alpha$-RMSD of 4.2 Å from the native state and the *SASA* increases by 8%, other changes relative to the native state are only marginal. For example, the radius of gyration decreases slightly, a result that may indicate that the $TSE_L$ has a more spherical structure than the native state (35). In comparison $TSE_H$ is much more unstructured, with an increase in *SASA* of 41% and an increase in *Rg* of 12%. In addition, clustering with a 3 Å threshold yielded 430 cluster centers whereas $TSE_L$ yielded only seven. The cluster centers of $TSE_L$ have very nativelike properties, whereas the $TSE_H$ cluster centers are very different in terms of C$\alpha$-RMSD, *Rg*, and *SASA* and are more heterogeneous, as judged by the significantly higher number of cluster centers (see Fig. 3).

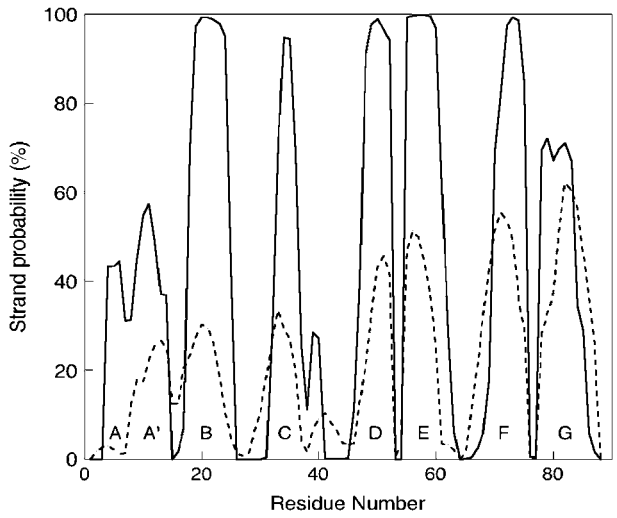


FIGURE 4 The secondary structure content of the TSEs. Probability of $\beta$-strand formation (in %) for $TSE_H$ (*dashed line*) and $TSE_L$ (*solid line*). The secondary structure was calculated using the program DSSPcont (78). The F and G strands appear to have the highest probability in $TSE_H$ but the results concerning the G strand may depend on the force field used because reliable experimental Φ-values in this strand could not be measured. Thus it is not possible to benchmark calculated Φ-values in this strand. The E and D strands are predicted to have a significant probability of formation, whereas the A, A′, and B strands are relatively unlikely to be formed. In comparison to $TSE_H$, in $TSE_L$ all strands are almost fully formed.

### *Interactions and secondary structure*

The secondary structure content of the $TSE_L$ has already been analyzed in detail (35). It is very nativelike and almost all strands are fully formed, except for the A, A′, and G strands. As well as the B, C, E, and F strands, the D strand, which is not part of the nucleus, is also fully structured. In the $TSE_H$ only the D, E, and F strands are significantly structured (see also Fig. 4).

In comparison to $TSE_H$, the interactions of $TSE_L$ are very nativelike (see Fig. 5). In $TSE_L$, most of the long-range interactions are defined as clearly as the short-range interactions in the energy map, as expected in a nucleation-condensation mechanism (35,43).

### *Compatibility between loosely structured individual conformers and the native topology indicated by the energy maps*

According to the analysis of the energy maps (Fig. 5), in the $TSE_H$, native interactions are still dominant over nonnative ones and the overall topology of the Ig-like fold is present. This result is in apparent contrast with a direct examination of individual structures. Although these structures show partial formation of secondary structure, they are relatively heterogeneous and contain many nonnative contacts. Our calculations show that all these observations are indeed consistent. First, individual structures are high-resolution descriptions of the protein whereas energy maps emphasize the general topology of the molecule. Second, whereas individual structures are snapshots of the TSE, energy maps are obtained from an averaging procedure over the whole ensemble and therefore they show the interactions that are most likely to be formed. Lindorff-Larsen et al. (30) showed that TS structures could be compared to structures in the SCOP database (44) to determine whether nativelike topology is already established in the TSE. Following this procedure, 770 protein domains of a length ranging from 71 to 110 residues were extracted from the SCOP database (44) to represent protein domains from a variety of folds. Using the DALI method (45) we classified the structures of the $TSE_H$. To determine if individual members of the $TSE_H$ can be classified as having Ig-like topology we aligned the 430 cluster centers that represent the $TSE_H$ to each of the 770 protein domains from the SCOP database (44). Analyzing the best hits from the alignments revealed that 61% of the cluster centers (which represent 55% of the total ensemble) can be classified as having the Ig-like fold (SCOP domain b.1). The quality of a DALI alignment (45) can be described using a Z-score; the higher the Z-score, the better the alignment and thus the more significant is the result. It has been suggested that a Z-score between 2 and 6 is the borderline for structural similarity between native-state structures (46). For the 61% of the structures for which the Ig-like fold was the best hit a Z-score between 1.0 and 5.0 was obtained, thus the similarity is universally low. In contrast all seven cluster centers and thus the whole $TSE_L$ ensemble match the Ig-like

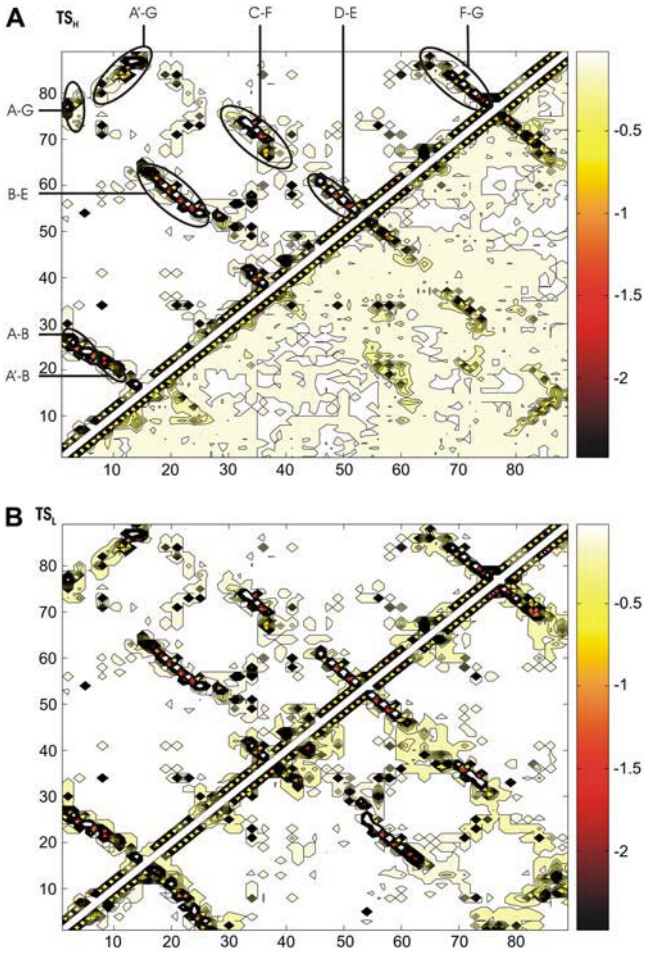


FIGURE 5 Interactions in the TSEs and equilibrated native states. Energy maps in the equilibrated native state (*upper part* of the *matrix*) and the transition state ensemble (*lower part* of the *matrix*) for (*A*) $TSE_H$ and (*B*) $TSE_L$. The energy map is a graphical representation of interaction matrices where the element *i*, *j* is the EEF1 (69) interaction energy between residues *i* and *j*, averaged in the TSE and in the equilibrated native state. Energies are in kcal/mol.

fold best, with Z-scores between 5.3 and 8.9. Hence they are also more similar to the Ig-fold than the 55% of $TSE_H$ that match best the Ig-like fold. Where a member of the $TSE_H$ could not be aligned to the topology of the Ig-like fold the structural similarity to any other ''best hit'' was very low (Z-score $< 1.0$); i.e., there was no significant match to any other fold. This means that, as we have seen in the energy maps, the overall topology of the Ig-fold is established to some extent in $TSE_H$. This is consistent with the current view that the native topology is already established in the transition state, e.g., (22,23,30,37,47–51).

### *Key residue simulations*

The folding nucleus of $TSE_L$ has recently been analyzed in detail (35). F21, I23, W34, H56, L58, V71, and F73 in the B, C, E, and F strands are key residues for folding along pathway L. The large hydrophobic W34 forms the center of the folding nucleus and the remaining six residues pack onto it. Consistent with these results, the topology of the native state can be established by using the Φ-values of every subset of four residues that included one residue from each strand as restraints (35). W34 is the most important residue in the nucleus of $TSE_L$; key residue simulations with all key residues except W34 are unable to reproduce $TSE_L$ accurately. Thus in $TS_L$ there is a well-defined folding nucleus.

In this work, four residues, L58, L60, C63, and M67, were shown to play a key role in folding via pathway H. It is

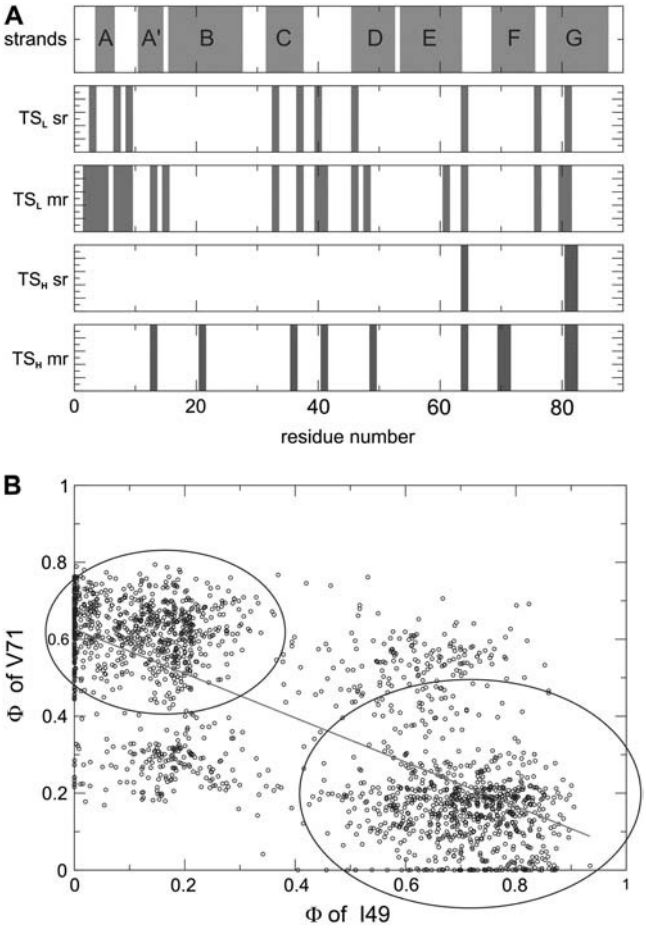


FIGURE 6 Distribution of Φ-value probabilities in the ensembles. (*A*) Probability distributions for Φ-values were calculated for single (sr) and multi- (mr) replica systems and for both TSE. The position of the strands is indicated in the first row. The second and third rows represent single and multireplica systems for $TS_L$, respectively. The fourth and fifth rows represent single and multireplica systems for $TS_H$, respectively. Residues with a nonunimodal distribution of probabilities for Φ-values are highlighted. All other residues have a unimodal distribution. Interestingly there are more residues in $TS_L$ with a nonunimodal distribution of probabilities for Φ-values than in $TS_H$. (*B*) $\Phi_{I49}$ and $\Phi_{V71}$ were calculated for each structure of the ensemble (two-replica system). They are anticorrelated with a correlation coefficient of $-0.75$; 80% of the structures are in one of the ellipsoids. This clearly shows that in most cases only I49 or V71 play an important role in folding.

important to note that the existence of key residues does not necessarily imply the existence of a folding nucleus. According to Vendruscolo et al. key residues are defined as residues whose interactions alone are sufficient to form the topology of the transition state and to predict all other Φ-values (3).

The Φ-value pattern (Fig. 1 *B*) is very different from the pattern expected for a classical nucleation-condensation mechanism, as found for example in $TS_L$ of TI I27 and another Ig-like protein TNfn3 (Fig. 1, *C* and *D*). $TSE_H$ is not dominated by long-range interactions but by local interactions, and most of the high Φ-values are colocalized. We determined the interactions and the distances between the key residues in the TSE. The average distance between key residue C$\beta$-atoms is 9 Å, and there are only four intrakey residue contacts. Thus we suggest that the protein does not fold via a conventional nucleation-condensation mechanism in pathway H.

We determined the packing density of the system of key residues for both TSEs by calculating the solvent-accessible surface area for each key residue (see Table 2). The solvent-accessible surface area is much lower for all the key residues in the $TSE_L$ (the packing density is much larger). Representative structures of both sets of key residues are shown in Fig. 7.

Interestingly, mutations that destabilize residues in the E-F loop, which is critical in the formation of structure in pathway H (C62A and M67A) fold only by pathway L under all experimental conditions whereas mutations that significantly destabilize the folding nucleus of pathway L (F21A, I23A, H56A, and F73L) reach pathway H at significantly lower concentrations of denaturant than wild-type. This supports the observations that are made here of the relative importance of these key residues in the different folding pathways.

### *The structure of the loop regions*

We have previously analyzed the behavior of the two loops crossing from one $\beta$-sheet to the other for $TSE_L$ in detail (35). The E-F loop connects the E strand with the F strand and the B-C loop connects the B strand with the C strand. In $TSE_L$ of TI I27 the nucleus lies closer to the B-C loop, therefore this loop is more structured. Thus, the relative position of the nucleus constrains the intersheet loops to different extents (52).

We studied the structural variability of the B-C and the E-F loops by isolating them from the rest of the protein and by clustering similar structures together using a 1-Å cutoff. In $TSE_L$ the average intraloop C$\alpha$-RMSD between two structures is 0.6 and 1.0 Å for the B-C and E-F loop, respectively, thus the B-C loop is more structured (35). In $TSE_H$ both loops are more unstructured but interestingly the E-F loop is now the more structured of the two. The average intraloop C$\alpha$-RMSD between two structures is 1.7 and 1.2 Å for the B-C and E-F loop, respectively. The structures of the loops are also more heterogeneous in $TSE_H$; the distributions are

**TABLE 2 Solvent-accessible surface area of key residues in the transition state**

| Transition state | Residue | $S$ (Å$^2$)* |
|---|---|---|
| $TS_H$† | L58 | 16 |
| | L60 | 18 |
| | C63 | 16 |
| | M67 | 34 |
| $TS_L$‡ | F21 | 1 |
| | I23 | 1 |
| | W34 | 6 |
| | H56 | 1 |
| | L58 | 1 |
| | V71 | 1 |
| | F73 | 1 |

*$S$ is the mean solvent-accessible surface area of a key residue.
†Key residues of $TS_H$.
‡Key residues of $TS_L$.

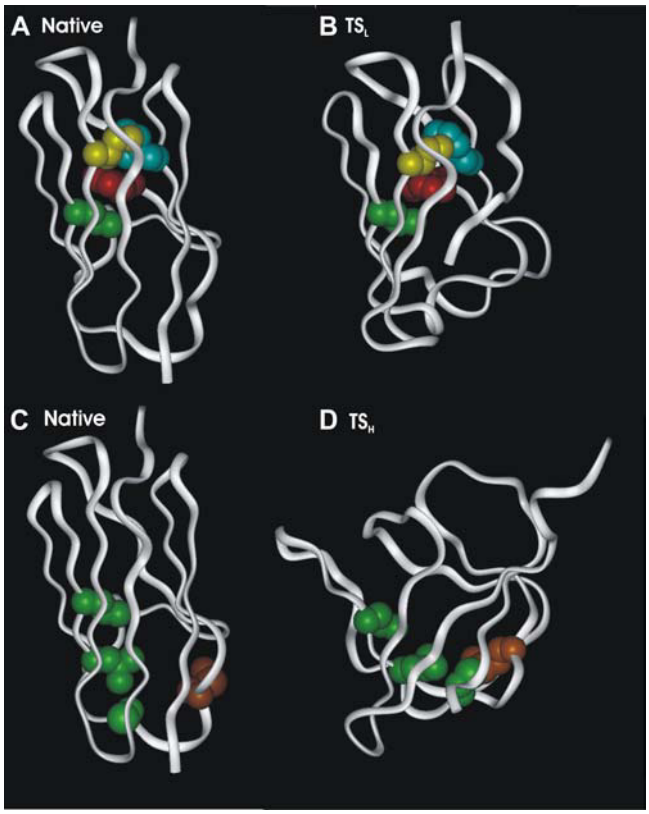


FIGURE 7 Comparison of the packing density in the folding cores of TI I27 for both pathways. The packing density is much higher in $TSE_L$ than in $TSE_H$, as illustrated by the CPK representations of the folding core of both transition states. Only the key residues are shown and they are colored according to the $\beta$-strand in which they are located. (*A*) Native state and (*B*) highest populated cluster center of the TSE of $TS_L$: I23 (*yellow*), W34 (*red*), L58 (*green*), and F73 (*blue*). For comparison only, one key residue from each strand is shown. (*C*) Native state and (*D*) highest populated cluster center of the TSE of $TS_H$: L58, (*green*), L60 (*green*), C63 (*green*), and M67 (*brown*). The pictures were generated using the program INSIGHTII.

much broader for this unstructured transition state (data not shown).

## Comparison of $TSE_H$ with the transition state of TNfn3

In the ''fold approach'' proteins with similar structures that are unrelated in sequence or function are analyzed. We have investigated the folding of a number of proteins with Ig-like topology extensively using Φ-value analysis (7,41,52,53). The TSE of TNfn3 and $TSE_L$ were determined using experimental Φ-values as restraints in molecular dynamics simulations and compared in detail (35,36). These studies revealed that both TNfn3 and TI I27 ($TS_L$) fold by a nucleation-condensation mechanism with structurally equivalent residues forming the folding nucleus. The determination of the second transition state of TI I27 enables us to compare $TS_H$ to that of TNfn3.

At a first glance, one might expect $TSE_H$ to be very similar to the TSE of TNfn3 in terms of overall structural features. TI I27 and TNfn3 have the same topology, i.e., the Ig-like fold, and about the same number of residues. The average experimental Φ-values are very similar (0.24 and 0.28 for $TS_H$ and TNfn3, respectively) so are the $\beta_T$ values (0.74 and 0.7 for $TS_H$ and TNfn3, respectively) (7,41,52). The patterns of Φ-values are, however, very different, because the Φ-value pattern of TNfn3 appears simply to be a weaker version of the $TSE_L$ Φ-value pattern, indicating a nucleation-condensation mechanism. In contrast $TSE_H$ is a polarized transition state.

The general properties of the TSE of TNfn3 as presented in Table 1 are very similar to those of $TSE_H$, consistent with the similar values of $<\Phi>$ and $\beta_T$. The change in *SASA* and *Rg* relative to the native state and the Cα-RMSD to the native state are remarkably similar. We also clustered the 4000 structures that we determined to represent the TSE of TNfn3 together using a 3-Å cutoff, obtaining 430 cluster centers. This result is identical to that obtained from clustering in the TSE of $TSE_H$, which also gave 430 clusters. A closer look, however, at the properties of the TSE of TNfn3 and $TSE_H$ confirms the suggestion that they are very different. Molecular dynamics simulations of TNfn3 restrained using the Φ-values of just the four residues I19, Y35, I58, and V69 suggested that these are the residues that interact in the core to form the folding nucleus (36). This nucleation-condensation transition state is stabilized mainly by long-range tertiary interactions. In $TSE_H$ the key residues L58, L60, C63, and M67 are localized, as expected in a polarized transition state, and do not interact to form a nucleus. Hence the distribution of structure is very different. Paci et al. showed that the C′-C-F-G sheet is largely ordered in the TSE of TNfn3, but the A-B-E sheet is disordered (36). Although the relative position of the B and E strands is established by interactions of residues I49 and I58 with the other nucleus residues, the hydrogen bonding between the two strands is minimal. This situation is very different for the secondary structure in $TSE_H$ where only the E-D strands are significantly ordered.

Thus although many of the characteristics of the TSE of TNfn3 and $TSE_H$ are similar (for example mean Φ-value, $\beta_T$, Δ*SASA*, RMSD), the structures in the ensembles are actually remarkably different. These results reflect the different nature of the transition states—a polarized TS in TI I27 and a diffuse TS in TNfn3. The TS of TNfn3 is mainly stabilized by long-range tertiary interactions whereas in the TS of pathway H of TI I27 long-range contacts are significantly weakened. This in turn suggests a different folding mechanism.

## Obligate versus critical residues in the folding nucleus

It has been suggested that the folding nucleus can be decomposed into obligate and critical residues (32,33,40). The former are obliged to participate with their interactions so that the folding nucleus can be formed and the topology can be established. These interactions are supported by the critical residues that stabilize the obligate nucleus by packing onto it. In both $TS_L$ and $TS_H$ we found several residues that participate in folding in only a subset of the TSE. Interestingly, the Φ-values of some of these residues are highly anticorrelated, meaning that only one of them supports folding and the other not. In $TS_H$ either I49 or V71 pack onto the nucleus to stabilize it and in $TS_L$ either 41 or 76. Thus for these residues there are two possible conformations for each transition state; suggesting the existence of ''parallel pathways in the parallel pathways''. In one of the earliest works describing Φ-value analysis Fersht and co-workers suggested that ''It is ... possible that a mixture of states, some with structure fully formed and others with the elements of structure completely unfolded is responsible for a fractional value of Φ.'' (54). This is what we observe in our simulations.

## Diffuse versus polarized transition states

It has been suggested that if there is a high propensity for secondary structure formation, the folding mechanism tends to follow a hierarchical mechanism, such as the diffusion-collision model, as has been shown for En-HD (20) and protein G (21,55). In contrast, if the propensity for formation of secondary structure is low, secondary structure elements will only form when supported by tertiary interactions. The resulting diffuse transition state is characteristic of the nucleation-condensation mechanism, where the formation of a well-defined nucleus is achieved simultaneously with the condensation of the rest of the structure resulting in a nativelike topology (22). This type of transition state shows a characteristic pattern of high and low fractional Φ-values distributed throughout the molecule (see, for example, Fig. 1, *C* and *D*); hence the term ''diffuse'' transition state. Such a

pattern often arises from the presence of a folding nucleus stabilized by long-range tertiary contacts (22,23,56). The majority of proteins that have been studied by Φ-value analysis fold via the nucleation-condensation mechanism, e.g., CI2 (23,57,58), the immunity proteins (59,60), and proteins with the Ig-like fold such as TNfn3 or TI I27 (35,36,41,52). In a classical nucleation-condensation mechanism, secondary and tertiary structure are formed concomitantly in the TS in the absence of intermediates, as found for example for hTRF (20). Thus, the transition state for folding in a nucleation-condensation pathway is often an expanded version of the native state (6,22). It is important to note that the term ''diffuse'' transition state does not necessarily imply a low density packed folding nucleus, but it rather describes a particular type of distribution of contacts. In this work we have shown that the folding nucleus of $TSE_L$ in TI I27 is indeed very compact and that it contains predominantly long-range contacts, and thus ''high'' Φ-values are distributed widely.

In contrast the polarized transition state seems to represent a hybrid between nucleation-condensation and diffusion-collision folding mechanisms (20,24,25). High Φ-values are not evenly distributed over the molecule but usually clustered at one position, with the rest of the structure exhibiting mainly low Φ-values; hence they are also called ''localized transition states'' (see Fig. 1 *B*). A polarized TSE is observed in SH3 domains, in which fully established contacts are mostly observed in turns that have to be formed to bring the remaining chain together (26–30). Several WW domains also seem to fold via a polarized transition state, whereby folding is initiated by loop or hairpin nucleation events (61–63). A polarized transition state is not necessarily unstructured: Garcia-Mira et al. recently examined the transition state for folding of CspB using Φ-value analysis (25). CspB folds via a strongly polarized transition state with a $\beta_T$ of 0.9, thus most regions of the protein are folded in the TS, but the energetic interactions are not yet properly established (25). The most polarized Φ-value distribution, and thus transition state, has been observed in the circular permutant *P*13-14 of S6 (6). This result is particularly interesting because wild-type S6 usually folds via a nucleation-condensation mechanism (31,32). In the circular permutant, the folding nucleus is polarized toward the artificially linked C- and N-termini.

Weikl et al. have used a theoretical approach based on the ''effective contact order'' to compute the rates and routes of folding of two-state proteins (64). In CI2, a protein with a typical nucleation-condensation mechanism (diffuse transition state) all major regions are involved in the formation of the rate-limiting cluster. In contrast the src-SH3 domain, which exhibits a polarized transition state, uses only regions with high Φ-values to form the rate-limiting cluster. Their results support the discrimination between diffuse and polarized transition states.

We propose that $TSE_H$ is a polarized transition state. Polarized transition states usually feature only local formation of secondary structure. The Φ-value pattern of $TSE_H$ has the typical profile of a polarized transition state; except for I49, all residues outside the region from 58 to 71 are very low. In this transition state, short-range local contacts are dominant. We found secondary structure mainly in the region of high Φ-values; most of the remaining regions in the protein are significantly unstructured. The structure distribution in $TSE_H$ can be categorized as being anisotropic compared to the isotropic distribution of structure in $TSE_L$.

## CONCLUSIONS

TI I27 can fold via two different pathways. The first is dominant in moderate conditions, such as those found in vivo, and exhibits a nucleation-condensation mechanism as found for most two-state proteins (diffuse TS). In contrast, TI I27 folds via a more polarized transition state at extreme values of denaturant or temperature. Highly destabilizing mutations in the nucleus of pathway L cause the folding mechanism to switch to pathway H. Conversely mutations of the key residues in the E-F loop prevents the switch to pathway H. The ensemble of structures that represent the two transition states for folding of TI I27 enable us to suggest an explanation for the change in mechanism. At near-physiological conditions, nucleation condensation seems to be preferred for proteins with Ig-like fold, despite the higher entropic cost for the protein chain to fold into the highly compact structure of $TSE_L$ (41,65). This is analogous to the observation from simulation that the folded state of a structured peptide is preferentially stabilized over low energy, high entropy alternative, nonnative conformations (66). Almost the entire protein chain is involved in the folding process, thus exhibiting a network of long-range tertiary interactions. This highly cooperative folding process ensures that that partly folded species with exposed hydrophobic residues do not form before crossing the top of the free-energy barrier. It has been suggested that this kind of folding mechanism, which involves burial of most hydrophobic groups, reflects the presence of a negative design feature against aggregation (4–8). The entropic cost of structuring the polypeptide chain is compensated for by burial of the hydrophobic residues. In contrast, at high concentration of denaturant, when the hydrophobic interactions are weakened, to fold efficiently TI I27 can revert to a mechanism with a lower entropic cost and a more heterogeneous transition state, because the propensity to aggregate is lower under these conditions.

C.D.G. holds a Wellcome Trust Prize Studentship. E.P. acknowledges financial support from Forschungskredit der Universität Zürich. M.V. is supported by the Royal Society and both M.V. and R.B. by the Leverhulme Trust. J.C. is a Wellcome Trust Senior Research Fellow.

## REFERENCES

1. Fersht, A. R., L. S. Itzhaki, N. Elmasry, J. M. Matthews, and D. E. Otzen. 1994. Single versus parallel pathways of protein-folding and fractional formation of structure in the transition-state. *Proc. Natl. Acad. Sci. USA*. 91:10426–10429.

2. Lazaridis, T., and M. Karplus. 1997. ''New view'' of protein folding reconciled with the old through multiple unfolding simulations. *Science*. 278:1928–1931.

3. Vendruscolo, M., E. Paci, C. M. Dobson, and M. Karplus. 2001. Three key residues form a critical contact network in a protein folding transition state. *Nature*. 409:641–645.

4. Davis, R., C. M. Dobson, and M. Vendruscolo. 2002. Determination of the structures of distinct transition state ensembles for a beta-sheet peptide with parallel folding pathways. *J. Chem. Phys*. 117:9510–9517.

5. Dobson, C. M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci*. 24:329–332.

6. Lindberg, M., J. Tangrot, and M. Oliveberg. 2002. Complete change of the protein folding transition state upon circular permutation. *Nat. Struct. Biol*. 9:818–822.

7. Wright, C. F., K. Lindorff-Larsen, L. G. Randles, and J. Clarke. 2003. Parallel protein-unfolding pathways revealed and mapped. *Nat. Struct. Biol*. 10:658–662.

8. Wright, C. F., A. Steward, and J. Clarke. 2004. Thermodynamic characterisation of two transition states along parallel protein folding pathways. *J. Mol. Biol*. 338:445–451.

9. Dinner, A. R., A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus. 2000. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci*. 25:331–339.

10. Wu, Y., and C. R. Matthews. 2003. Proline replacements and the simplification of the complex, parallel channel folding mechanism for the alpha subunit of Trp synthase, a TIM barrel protein. *J. Mol. Biol*. 330:1131–1144.

11. Wedemeyer, W. J., E. Welker, and H. A. Scheraga. 2002. Proline cis-trans isomerization and protein folding. *Biochemistry*. 41:14637–14644.

12. Wildegger, G., and T. Kiefhaber. 1997. Three-state model for lysozyme folding: triangular folding mechanism with an energetically trapped intermediate. *J. Mol. Biol*. 270:294–304.

13. Zaidi, F. N., U. Nath, and J. B. Udgaonkar. 1997. Multiple intermediates and transition states during protein unfolding. *Nat. Struct. Biol*. 4:1016–1024.

14. Shastry, M. C., and J. B. Udgaonkar. 1995. The folding mechanism of barstar: evidence for multiple pathways and multiple intermediates. *J. Mol. Biol*. 247:1013–1027.

15. Baldwin, R. L. 1997. Competing unfolding pathways. *Nat. Struct. Biol*. 4:965–966.

16. Honeycutt, J. D., and D. Thirumalai. 1990. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci. USA*. 87:3526–3529.

17. Shimada, J., and E. I. Shakhnovich. 2002. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci. USA*. 99:11175–11180.

18. Ferrara, P., and A. Caflisch. 2000. Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc. Natl. Acad. Sci. USA*. 97: 10780–10785.

19. Nauli, S., B. Kuhlman, and D. Baker. 2001. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol*. 8:602–605.

20. Gianni, S., N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. White, M. L. DeMarco, V. Daggett, and A. R. Fersht. 2003. Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA*. 100:13286–13291.

21. Islam, S. A., M. Karplus, and D. L. Weaver. 2004. The role of sequence and structure in protein folding kinetics; the diffusion-collision model applied to proteins L and G. *Structure*. 12:1833–1845.

22. Fersht, A. R. 1995. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA*. 92:10869–10873.

23. Itzhaki, L. S., D. E. Otzen, and A. R. Fersht. 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol*. 254:260–288.

24. Grantcharova, V., E. J. Alm, D. Baker, and A. L. Horwich. 2001. Mechanisms of protein folding. *Curr. Opin. Struct. Biol*. 11:70–82.

25. Garcia-Mira, M. M., D. Boehringer, and F. X. Schmid. 2004. The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol*. 339:555–569.

26. Grantcharova, V. P., D. S. Riddle, J. V. Santiago, and D. Baker. 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Biol*. 5: 714–720.

27. Riddle, D. S., V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker. 1999. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol*. 6:1016–1024.

28. Shea, J. E., J. N. Onuchic, and C. L. Brooks 3rd. 2002. Probing the folding free energy landscape of the Src-SH3 protein domain. *Proc. Natl. Acad. Sci. USA*. 99:16064–16068.

29. Guo, W., S. Lampoudi, and J. E. Shea. 2004. Temperature dependence of the free energy landscape of the src-SH3 protein domain. *Proteins*. 55:395–406.

30. Lindorff-Larsen, K., M. Vendruscolo, E. Paci, and C. M. Dobson. 2004. Transition states for protein folding have native topologies despite high structural variability. *Nat. Struct. Biol*. 11:443–449.

31. Otzen, D. E., and M. Oliveberg. 2002. Conformational plasticity in folding of the split beta-alpha-beta protein S6: evidence for burst-phase disruption of the native state. *J. Mol. Biol*. 317:613–627.

32. Hubner, I. A., M. Oliveberg, and E. I. Shakhnovich. 2004. Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. *Proc. Natl. Acad. Sci. USA*. 101:8354–8359.

33. Krantz, B. A., R. S. Dothager, and T. R. Sosnick. 2004. Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol*. 337:463–475.

34. Sanchez, I. E., and T. Kiefhaber. 2003. Origin of unusual phi-values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol*. 334:1077–1085.

35. Geierhaas, C. D., E. Paci, M. Vendruscolo, and J. Clarke. 2004. Comparison of the transition states for folding of two Ig-like proteins from different superfamilies. *J. Mol. Biol*. 343:1111–1123.

36. Paci, E., J. Clarke, A. Steward, M. Vendruscolo, and M. Karplus. 2003. Self-consistent determination of the transition state for protein folding: application to a fibronectin type III domain. *Proc. Natl. Acad. Sci. USA*. 100:394–399.

37. Paci, E., K. Lindorff-Larsen, C. M. Dobson, M. Karplus, and M. Vendruscolo. 2005. Transition state contact orders correlate with protein folding rates. *J. Mol. Biol*. 352:495–500.

38. Paci, E., M. Vendruscolo, C. M. Dobson, and M. Karplus. 2002. Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol*. 324:151–163.

39. Vendruscolo, M., N. V. Dokholyan, E. Paci, and M. Karplus. 2002. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E*. 65:061910.

40. Hedberg, L., and M. Oliveberg. 2004. Scattered Hammond plots reveal second level of site-specific information in protein folding: phi′ (beta++). *Proc. Natl. Acad. Sci. USA*. 101:7606–7611.

41. Fowler, S. B., and J. Clarke. 2001. Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. *Structure*. 9:355–366.

42. Salvatella, X., C. M. Dobson, A. R. Fersht, and M. Vendruscolo. 2005. Determination of the folding transition states of barnase by using PhiI-value-restrained simulations validated by double mutant PhiIJ-values. *Proc. Natl. Acad. Sci. USA*. 102:12389–12394.

43. Fersht, A. R. 1997. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol*. 7:3–9.

44. Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. 247:536–540.

45. Holm, L., and C. Sander. 1996. Mapping the protein universe. *Science*. 273:595–603.

46. Dietmann, S., N. Fernandez-Fuentes, and L. Holm. 2002. Automated detection of remote homology. *Curr. Opin. Struct. Biol.* 12:362–367.

47. Baker, D. 2000. A surprising simplicity to protein folding. *Nature*. 405: 39–42.

48. Otzen, D. E., L. S. Itzhaki, N. F. elMasry, S. E. Jackson, and A. R. Fersht. 1994. Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc. Natl. Acad. Sci. USA.* 91:10422–10425.

49. Plaxco, K. W., K. T. Simons, and D. Baker. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

50. Fersht, A. R. 2000. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. USA*. 97: 1525–1529.

51. Ferrara, P., and A. Caflisch. 2001. Native topology or specific interactions: what is more important for protein folding? *J. Mol. Biol.* 306:837–850.

52. Hamill, S. J., A. Steward, and J. Clarke. 2000. The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* 297: 165–178.

53. Cota, E., and J. Clarke. 2000. Folding of beta-sandwich proteins: three-state transition of a fibronectin type III module. *Protein Sci.* 9:112–120.

54. Fersht, A. R., A. Matouschek, and L. Serrano. 1992. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* 224:771–782.

55. McCallister, E. L., E. Alm, and D. Baker. 2000. Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.* 7:669–673.

56. Ternstrom, T., U. Mayor, M. Akke, and M. Oliveberg. 1999. From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. USA*. 96:14854–14859.

57. Li, A. J., and V. Daggett. 1996. Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* 257:412–429.

58. Li, L., and E. I. Shakhnovich. 2001. Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA*. 98:13014–13018.

59. Friel, C. T., A. P. Capaldi, and S. E. Radford. 2003. 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.* 326:293–305.

60. Paci, E., C. T. Friel, K. Lindorff-Larsen, S. E. Radford, M. Karplus, and M. Vendruscolo. 2004. Comparison of the transition state ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with phi value restraints. *Proteins.* 15:513–525.

61. Jager, M., H. Nguyen, J. C. Crane, J. W. Kelly, and M. Gruebele. 2001. The folding mechanism of a beta-sheet: the WW domain. *J. Mol. Biol.* 311:373–393.

62. Ferguson, N., and A. R. Fersht. 2003. Early events in protein folding. *Curr. Opin. Struct. Biol.* 13:75–81.

63. Ferguson, N., J. R. Pires, F. Toepert, C. M. Johnson, Y. P. Pan, R. Volkmer-Engert, J. Schneider-Mergener, V. Daggett, H. Oschkinat, and A. Fersht. 2001. Using flexible loop mimetics to extend phi-value analysis to secondary structure interactions. *Proc. Natl. Acad. Sci. USA*. 98:13008–13013.

64. Weikl, T. R., and K. A. Dill. 2003. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* 329:585–598.

65. Hamill, S. J., E. Cota, C. Chothia, and J. Clarke. 2000. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. *J. Mol. Biol.* 295:641–649.

66. Rao, F., and A. Caflisch. 2004. The protein folding network. *J. Mol. Biol.* 342:299–306.

67. Improta, S., A. S. Politou, and A. Pastore. 1996. Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity. *Structure.* 4:323–337.

68. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

69. Lazaridis, T., and M. Karplus. 1999. Effective energy function for proteins in solution. *Proteins.* 35:133–152.

70. Paci, E., and M. Karplus. 1999. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.* 288:441–459.

71. Best, R. B., and M. Vendruscolo. 2004. Determination of protein structures consistent with NMR order parameters. *J. Am. Chem. Soc.* 126:8090–8091.

72. Best, R. B., and M. Vendruscolo. 2006. Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Structure.* 14:97–106.

73. Paci, E., M. Vendruscolo, and M. Karplus. 2002. Validity of Go models: comparison with a solvent-shielded empirical energy decomposition. *Biophys. J.* 83:3032–3038.

74. Sayle, R. A., and E. J. Milnerwhite. 1995. Rasmol: biomolecular graphics for all. *Trends Biochem. Sci.* 20:374–376.

75. Kraulis, P. J. 1991. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24:946–950.

76. Merritt, E. A., and D. J. Bacon. 1997. Raster3D: photorealistic molecular graphics, in macromolecular crystallography, Part B. *Methods Enzymol.* 277:505–524.

77. Koradi, R., M. Billeter, and K. Wuthrich. 1996. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* 14:29–32.

78. Carter, P., C. A. F. Andersen, and B. Rost. 2003. DSSPcont: continuous secondary structure assignments for proteins. *Nucleic Acids Res.* 31:3293–3295.