Frustration and folding of a TIM barrel protein

Kevin T. Halloran[a,1], Yanming Wang[b,1], Karunesh Arora[b], Srinivas Chakravarthy[c], Thomas C. Irving[c], Osman Bilsel[a], Charles L. Brooks III[b,2], and C. Robert Matthews[a,2]

[a]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605; [b]Department of Chemistry and Biophysics Program, University of Michigan, Ann Arbor, MI 48109; and [c]Biophysics Collaborative Access Team (BioCAT), Center for Synchrotron Radiation Research and Instrumentation and Department of Biological and Chemical Sciences, Illinois Institute of Technology, Chicago, IL 60616

Edited by William F. DeGrado, University of California, San Francisco, CA, and approved June 27, 2019 (received for review January 22, 2019)

Triosephosphate isomerase (TIM) barrel proteins have not only a conserved architecture that supports a myriad of enzymatic functions, but also a conserved folding mechanism that involves on- and off-pathway intermediates. Although experiments have proven to be invaluable in defining the folding free-energy surface, they provide only a limited understanding of the structures of the partially folded states that appear during folding. Coarse-grained simulations employing native centric models are capable of sampling the entire energy landscape of TIM barrels and offer the possibility of a molecular-level understanding of the readout from sequence to structure. We have combined sequence-sensitive native centric simulations with small-angle X-ray scattering and time-resolved Förster resonance energy transfer to monitor the formation of structure in an intermediate in the *Sulfolobus solfataricus* indole-3-glycerol phosphate synthase TIM barrel that appears within 50 μs and must at least partially unfold to achieve productive folding. Simulations reveal the presence of a major and 2 minor folding channels not detected in experiments. Frustration in folding, i.e., backtracking in native contacts, is observed in the major channel at the initial stage of folding, as well as late in folding in a minor channel before the appearance of the native conformation. Similarities in global and pairwise dimensions of the early intermediate, the formation of structure in the central region that spreads progressively toward each terminus, and a similar rate-limiting step in the closing of the β-barrel underscore the value of combining simulation and experiment to unravel complex folding mechanisms at the molecular level.

protein-folding intermediates | Gō models | TIM barrel protein

The folding of globular proteins involves the formation of hundreds of noncovalent interactions as the polypeptide chain samples the folding free energy surface on its journey to the global free energy minimum. Given the complexity of the conformational transition, it is surprising that proteins execute their folding reactions within a few 10s of seconds on a relatively smooth energy surface (1). The folding reactions of small proteins and domains, <100 amino acids, usually follow the expectations for a 2-state process with a single barrier between unfolded and native states that controls a simple exponential response. Larger proteins, however, often have more complex responses involving multiple exponential phases, the rate constants of which progressively decrease as they approach the native state (2). These phases have been attributed to the appearance of partially folded states, the role of which in folding may be to avoid aggregation, accelerate folding, or simply be a consequence of the interplay between the sequence and topology of the protein (3). A complementary view of such intermediates posits that their presence reflects frustration in folding that precludes the direct formation of the native state by the topological incompatibility of preformed elements of structure (4, 5). Experiments and simulations have revealed that these intermediates may be native-like in secondary structure but contain nonnative interactions (6) or contain structural elements not found in the native structure (7).

The triosephosphate isomerase (TIM) barrel family (Fig. 1*A*) provides a rich example of complex folding reactions, the kinetic responses of which are largely conserved while the underlying sequences vary widely (8). Previous folding studies on several family members have consistently found that folding comprised a submillisecond burst phase followed by a phase with a relaxation time that decreases with increasing final urea concentrations, a characteristic of an unfolding reaction but under refolding conditions. The escape from a misfolded, off-pathway intermediate is followed by the further acquisition of secondary structure and stability in an on-pathway intermediate(s) and the rate-limiting formation of the native state (8, 9). The kinetically trapped species could be frustration in folding whereby the burst phase species must unfold to enable productive folding to the native state. Mutational analysis and hydrogen exchange (HDX) experiments on several TIM barrel proteins have revealed a relationship between sequence and topology in the structures of the kinetically trapped and on-pathway intermediates (8, 10, 11). Clusters of branched aliphatic side chains, isoleucine, leucine, and valine, local in sequence and local in space, form water-resistant cores that stabilize these partially folded forms (12). Although the kinetic species are conserved, the evolution of the sequences over time has resulted in alternative locations for the cores of stability in the TIM barrel architecture.

The mutational and HDX experiments are valuable in pinpointing the regions where structure appears in intermediates detected after the first few milliseconds of folding but leave unanswered questions about the crucial initial folding reaction. Recent advances in microfluidic mixers now enable access to

Significance

The ways in which proteins fold to their native state remain a challenging and important question in biology. Large proteins often populate meta-stable intermediate states before reaching their native conformation, making them excellent targets for understanding the sequence-structure code. We studied folding of indole-3-glycerol phosphate synthase, a $(\beta\alpha)_8$ TIM barrel protein, using SAXS and FRET experiments and coarse-grained computer simulation to gain insights into structures of early folding intermediates. The folding mechanism first proposed by experimental studies is elaborated by a more complete mechanism involving 3 folding channels revealed by simulation. We provide connections between the off-pathway intermediate and one on-pathway rate-determining intermediate detected by experiment with corresponding meta-stable states in the major folding channel found in the simulations.

Author contributions: K.T.H., Y.W., T.C.I., O.B., C.L.B., and C.R.M. designed research; K.T.H., Y.W., K.A., S.C., and O.B. performed research; Y.W. and O.B. contributed new reagents/analytic tools; K.T.H., Y.W., S.C., O.B., C.L.B., and C.R.M. analyzed data; and K.T.H., Y.W., O.B., C.L.B., and C.R.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]K.T.H. and Y.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: brookscl@umich.edu or c.robert.matthews@umassmed.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1900880116/-/DCSupplemental.

Published online July 25, 2019.

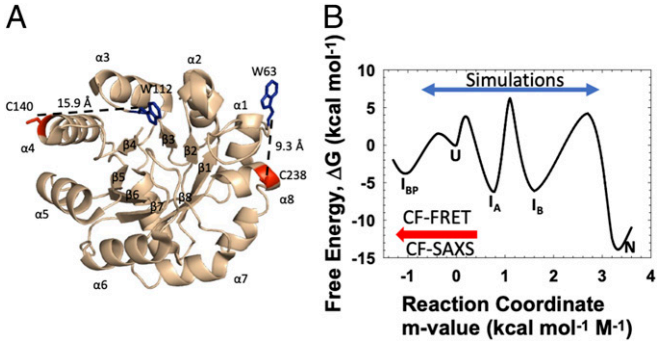

Fig. 1. (*A*) A ribbon representation of the structure of SsIGPS (PDB code 2C3Z) prepared by PyMol. The FRET pairs employed to study the central α_3-α_4 segment, W112-C140, [containing the strongly protected $(\beta\alpha)_{3-4}$ module] and the N- and C-termini, W63-C238, are highlighted with W112 and W63 residues in blue and the C140 and C238 residues in red. (*B*) The reaction diagram of SsIGPS (14). The barrier heights were estimated using Kramer's formalism with a prefactor of 1 μs. The blue arrows indicate aspects of the free energy landscape probed in this study by simulations, and the red arrow indicates the focus of the present experiments.

folding events in the microsecond time range and have allowed us to examine the earliest events in the folding of a candidate TIM barrel protein (13). In the present study, pair-wise distance measurements from time-resolved Förster resonance energy transfer (trFRET) and global size and shape measurements from small-angle X-ray scattering (SAXS), combined with coarse-grained computer simulations, are employed to probe the earliest events in the folding of the *Sulfolobus solfataricus* indole-3-glycerol phosphate synthase (SsIGPS) TIM barrel. Surprisingly, the global collapse of the unfolded chain to a misfolded, off-pathway intermediate occurs within 50 μs. The simulations reveal the potential complexities of this exceedingly rapid reaction and show that the rate-limiting step in the folding of SsIGPS is the frustration encountered by the competition between the N- and C-terminal β-strands to close the 8-stranded β-barrel.

Results

The SsIGPS TIM barrel is a representative member of the most common architecture for enzymes in biology. The 8 alternating β and α elements are arranged sequentially as a central parallel-stranded β-barrel encompassed by an α-helical shell (Fig. 1*A*). Its folding mechanism has previously been shown to begin with the submillisecond formation of an off-pathway intermediate, I_{BP}, followed by 2 on-pathway intermediates, I_A and I_B, before reaching the native state (Fig. 1*B*). The I_{BP} intermediate has an apparent stability of 3.5 kcal·mol^{-1}, is rich in secondary structure, and displays strong protection against exchange of amide hydrogens with solvent in the central $(\beta\alpha)_4$ module within 75 ms (14). As the folding reaction proceeds, the protection expands to encompass $(\beta\alpha)_{2-6}$ in I_A and $(\beta\alpha)_{1-8}$ in I_B. The fully folded TIM barrel appears in the final step of folding.

Measuring Global Dimensions by SAXS. To obtain global insights into the structures of the intermediates, SAXS profiles were obtained under equilibrium and kinetic refolding conditions. At equilibrium, the native state of the protein has a radius of gyration (Rg) of ~18 Å, and the unfolded state has an estimated Rg of 46 Å by linear extrapolation from high denaturant conditions (Fig. 2*A*). The I_A intermediate, highly populated at 4 M urea (14), self-associates at the 80-μM protein concentration required for reliable SAXS measurements, precluding an estimate of its Rg (*SI Appendix*, Fig. S1).

The Rg of the submillisecond burst-phase intermediate I_{BP} was determined by a 10-fold dilution from 8 M urea, using a custom, single-piece microfluidic mixer. Within 150 μs, the dead time of the mixer, the Rg of SsIGPS is 26 ± 1.5 Å (Fig. 2*B*). The absence of change in Rg out to 4 ms, where I_{BP} can be detected by stop-flow circular dichroism, shows that the I_{BP} intermediate appears within 150 μs. The conclusion that the SAXS-detected burst-phase species is a discrete thermodynamic state, and not a collapsed form of the unfolded state, is supported by the observation of a Rg that is insensitive to the urea concentration up to 2 M urea (Fig. 2*A*). A collapsed form of the unfolded state would have been expected to swell with increasing urea concentration (15).

Transformation of the scattering curve from the native state of SsIGPS in 0.8 M urea to a dimensionless Kratky plot (16) shows the parabolic shape typical of globular structure. The maximum in the plot occurs at $(\sqrt{3}, 1.1)$ as expected for the Guinier approximation (Fig. 2*C*) (16). At 8 M urea, SsIGPS has an extended random coil-like structure with the expected hyperbolic plateau shape at high scattering angle × Rg (qRg). By contrast, the dimensionless Kratky plot for the continuous flow refolding curve shows that the I_{BP} state has a peak shift on the qRg axis to approximately (2, 1.25). This behavior deviates from the Guinier approximation and shows that the protein has regions that are not yet fully globular. The pair distribution function, P(r), for I_{BP} confirms a large collapse of the chain from the unfolded state (U) to I_{BP} within 150 μs (Fig. 2*D*). The maximum distance between any 2 atoms, D_{max}, concomitantly decreases from 130 to 80 Å, and the significant shoulder at ~70 Å shows that I_{BP} is not fully globular.

Pair-Wise Dimensional Analysis by trFRET. To complement the global dimensional data obtained by SAXS, 2 sets of pair-wise distances were measured by trFRET experiments on SsIGPS, using tryptophan and AEDANS. One FRET pair was positioned to monitor barrel closure by measuring the distance distribution between α_1 and α_8, W63-C/AEDANS238. The second pair was positioned to monitor the formation of the strongly protected

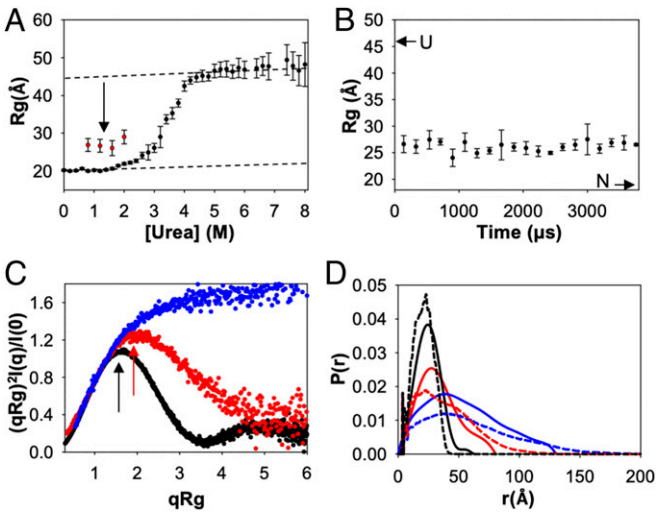

Fig. 2. (*A*) Rg as a function of urea concentration for the unfolding of SsIGPS (black circles). The estimated Rgs of the native and unfolded states in water are indicated by linear extrapolation of the native and unfolded baselines. The estimated Rg after 150 μs of refolding at several final urea concentrations in the native baseline region (red circles). (*B*) Rg as a function of folding time at 0.8 M urea. The Rgs of the unfolded and native states in water are indicated. (*C*) Dimensionless Kratky plots of the unfolded (blue), I_{BP} intermediate (red), and native state (black). The arrows indicate the maxima in the plots for the N and I_{BP} species. (*D*) The P(r) of the unfolded (blue), I_{BP} intermediate (red), and native states (black). The dashed lines represent the P(r) for these states calculated from the simulations.

$(\beta\alpha)_4$ module (8) by measuring the distance between α_3 and α_4, W112 and C/AEDANS140.

The average Trp lifetimes for donor-only (DO) and donor-acceptor (DA) samples for both the α_1-α_8 and α_3-α_4 pairs at 8 M urea were identical, consistent with the absence of FRET in the unfolded state (*SI Appendix*, Fig. S2 *A* and *B*). The continuous flow trFRET (CF-trFRET) data for the refolding of SsIGPS containing the α_1-α_8 pair shows a decrease to a non–native-like lifetime of 4.5 ns for the DO sample and 3.3 ns for the DA sample within the dead time of the mixer (50 μs) (*SI Appendix*, Fig. S2 *A* and *B*). As was the case for Rg, there are no significant changes in lifetimes for both the DO and DA samples from ∼50 μs out to ∼1 ms. Similar behavior was observed for the α_3-α_4 FRET pair during refolding jumps to 0.8 M urea, demonstrating a global collapse of unfolded SsIGPS within 50 μs.

Maximum Entropy Modeling. The trFRET data for both the α_1-α_8 and α_3-α_4 pairs were analyzed by 2D maximum entropy modeling (2D-MEM) (17, 18) for the unfolded state (8 M urea), I_{BP} (0.8 M urea, 100 μs), and the native state (0 M urea) (Fig. 3 *A* and *B* and *SI Appendix*, Fig. S2*C*). The unfolded state for both FRET pairs show very small normalized amplitudes from 12 to 35 Å, the distances most sensitive to FRET for the Trp-AEDANS pair (R_0 = 22 Å). The maximum normalized amplitude in the native state for the α_1-α_8 pair (Fig. 3*A*) is ∼13 Å, in good agreement with the calculated distance between residues 63 and 238 in the crystal structure, 9.3 Å. The maximum normalized amplitude in the native state for the α_3-α_4 FRET pair has a peak at ∼19 Å, also in good agreement with the distance between Cαs of residues 112 and 140 in the crystal structure, 15.9 Å. Surprisingly, the normalized amplitude after 100 μs for the α_1-α_8 FRET pair revealed the presence of 2 distinct distributions of distances. One distribution of distances is more compact than native, and the other is more extended than native but more compact than the unfolded state. By contrast, the α_3-α_4 pair after 100 μs shows a single peak around 20 Å (Fig. 3*B*) that is similar to the native protein but has a broader distribution.

Ensemble Averaged Folding Properties from Simulations. To gain deeper insight into the development of structure during the folding of SsIGPS, we complemented the present and previous experimental studies (8, 19) with a native centric simulation to sample the entire folding landscape.

Native-centric coarse-grained Gō model (20) refolding simulations were initiated from an unfolded ensemble of structures sampled from simulations at high temperature (see *Materials and Methods* for further details), and 100 independent 2,000 time units (1 time unit = 10,000 dynamics steps) folding trajectories were sampled in the analysis. Because the underlying model is coarse-grained, the landscape is smoother and the folding timescales are compressed and thereby do not directly correspond to the times observed in experiments. However, we anticipate that the time ordering, as well as the relative lag times between folding phases, should reflect what is observed in kinetic experiments (21). The progress of the folding reaction for each trajectory was monitored by the Rg and the fraction of total native contacts (Q_t). Over the time courses of the 100 trajectories, persistent values were observed for Q_t of 0.3, ∼0.5, ∼0.6, ∼0.8, and 0.9 (Fig. 4*A*). The initial (0.3) and final (0.9) values correspond to the unfolded and native forms of the protein with the intermediate plateaus (∼0.5, ∼0.6, ∼0.8) suggesting the presence of partially folded states. Examination of the entire set of trajectories revealed that only a small fraction of the simulations reached the native state within 2,000 time units (*SI Appendix*, Fig. S3). The majority of the simulations reached a Q_t of 0.8. Plateaus at similar time steps were observed for Rg (Fig. 4*B*), beginning with the unfolded state at ∼45 Å and progressively decreasing to 18 Å for the native state.

Decomposition of Q_t into contributions (Q_i) from the N- and C-terminal halves, $\alpha_0(\beta\alpha)_{1-4}$ and $(\beta\alpha)_{5-8}$, reveals frustration in folding (Fig. 4*C*). The striking anticorrelated gain/loss in native contacts in N- and C-terminal halves was observed at Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85, where the intermediate states persist, which suggests that the frustration in these 2 regions might be related to those intermediate states. Further decomposition of Q_t into the four $(\beta\alpha)_{i-(i+1)}$ modules of stability (22) pinpoint the major sources of frustration (Fig. 4*D*). At Q_t = 0.5 the source of frustration derives from $(\beta\alpha)_{1-2}$ competing with $(\beta\alpha)_{7-8}$ and to a lesser extent $(\beta\alpha)_{5-6}$. The frustration event at Q_t = 0.6 is $(\beta\alpha)_{1-2}$ driving folding while $(\beta\alpha)_{3-4}$ and $(\beta\alpha)_{5-6}$ lose native contacts. The final frustration event at Q_t = 0.8 is mainly the competition between the $(\beta\alpha)_{1-2}$ and $(\beta\alpha)_{7-8}$ modules; however, the competition also effects $(\beta\alpha)_{3-4}$ and $(\beta\alpha)_{5-6}$.

An examination of the contact probability maps at different times gives further insight into the folding mechanism (*SI Appendix*, Fig. S4). The central region (residue ∼90 to residue ∼180) formed most of its contacts within 400 time units compared with the total simulation time of 2,000 time units. Then, more contacts were formed in the C-terminal region within 1,000 time units. At the end of the simulation, most of the low-probability contacts were those formed between $\alpha_0\beta_1$ (residues ∼1 to ∼40) and other regions, suggesting that many trajectories ended up with a structure with unfolded $\alpha_0\beta_1$.

Multiple Folding Pathways Revealed from Simulations. To obtain further insights into the molecular events that occur during the folding of SsIGPS, we examined each trajectory in detail. Three significant folding pathways were found, based on the assembly order of secondary structural units (Fig. 5). The classification of different states found in all trajectories was based upon both their distinct Q_t and Rg values (*SI Appendix*, Table S1) and their visual differences in structure (Fig. 5 and *SI Appendix*, Fig. S5).

The U state initially collapses to a single intermediate, I_c, with a well-folded central region [residues 75–175, $(\beta\alpha)_{2-5}$]. The I_c state then partitions into the I_{1A} state, the I_2 state, or the I_3 state with transition probabilities of 81, 9, and 10%, respectively, entering 3 separate folding channels.

The I_1 pathway is characterized by the formation of an extremely stable 1_{1A} state after I_c. The I_c state transited to I_{1A} by spreading its folded structure from $(\beta\alpha)_{2-5}$ to $(\beta\alpha)_{2-8}$, leaving an unfolded α_0 tethered by β_1. The I_{1A} state has a 7-stranded barrel with native-like contacts between helices α_1 and α_8 that prevent the incorporation of α_0 and β_1 into the barrel architecture. I_{1A} persists in the great majority of the trajectories with only a small fraction escaping to dock α_0 across the bottom of the barrel to form the I_{1B} state. The I_{1B} state has an unfolded β_1 with 2 of its ends fixed on the folded β-barrel. I_{1B} then rapidly folds to the native state by the insertion of β_1 into the β-barrel through the

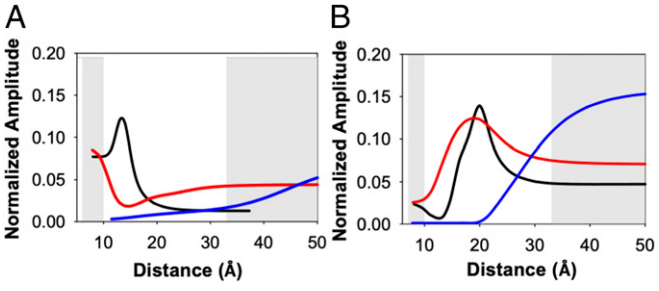

Fig. 3. (*A*) The 1D conversion of the MEM analysis of the trFRET data for the α_1-α_8 pair to generate distance distributions for the unfolded state (blue), the native state (black), and the apparent pair of states appearing after 50 μs (red). (*B*) The 1D conversion of the MEM analysis of the trFRET data for the α_3-α_4 pair to generate distance distributions for the unfolded state (blue), the native state (black), and the 2 states appearing after 50 μs (red).

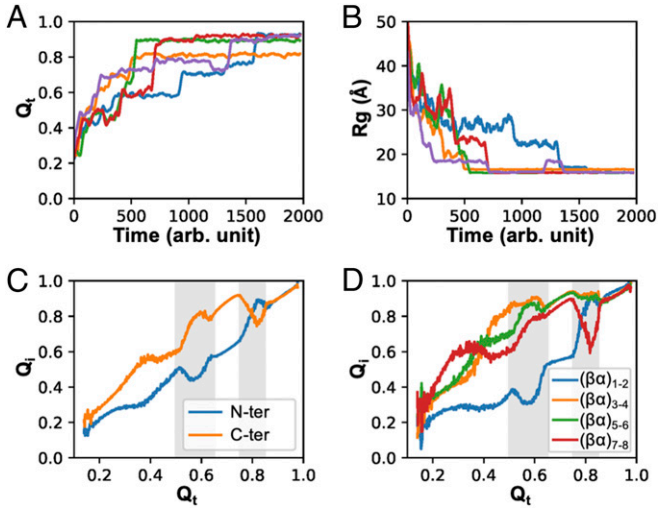

Fig. 4. (*A*) Representative trajectories of the fraction of total native contacts Q_t. (*B*) Representative trajectories of Rg. For clarity, kinetic traces are shown as moving averages of 30 successive snapshots. The leveling off at various Rg and Q_t values indicates that multiple intermediates are formed during the simulations. (*C*) Fraction of native contacts Q_i of the N- and C-terminal halves of the protein and (*D*) Q_i of the 4 βαβα modules as a function of Q_t. Decreases in the fraction of native contacts in *C* and *D* represent the backtracking that occurs during the simulations. There are 2 main backtracking events that involve the N- and C-termini (Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85).

channel between α_1 and α_8. To test the stability of the I_{1A} state, we further sampled another set of 100 trajectories beginning from the I_{1A} state for 8,000 time units. Even with a quadrupled simulation time, only 17% of the trajectories reached the native state (*SI Appendix*, Fig. S7), confirming the extremely long lifetime of the I_{1A} state.

In comparison with the I_1 pathway, in which $\alpha_0\beta_1$ is the last to fold, both the I_2 and the I_3 pathway require the $\alpha_0\beta_1$ element to fold before the closure of the β-barrel. In the I_2 pathway, I_c incorporates the $\alpha_0\beta_1$ element but excludes the C-terminal $(\alpha\beta)_{7-8}$ elements to form the I_2 state. I_2 then readily folds to the native state by incorporating the C-terminal $(\alpha\beta)_{7-8}$. In the I_3 pathway, I_c first forms 2 partially folded $\alpha_2(\alpha\beta)_{3-5}$ and $\alpha_6(\alpha\beta)_{7-8}$ before the docking of $\alpha_0\beta_1$ on $\alpha_6(\alpha\beta)_{7-8}$ to form the I_3 state. The I_3 state has 2 partially folded halves of the β-barrel, the $\alpha_2(\alpha\beta)_{3-5}$ subdomain, and the $(\alpha\beta)_0+\alpha_6(\alpha\beta)_{7-8}$ subdomain, linked by unfolded $\beta_1\alpha_1\beta_2$ and β_6 strands. Interestingly, the I_3 state, fully connected by contacts from head to tail, then folds to native by cooperatively merging the 2 partially folded halves. More structural details of the I_{1A}, I_{1B}, I_2, and I_3 intermediates are shown in *SI Appendix*, Fig. S5.

To examine the relationship between the 2 regions of frustration (Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85) found in the ensemble averaged analysis (Fig. 4*C*) and the 3 folding pathways, the Q_i vs. Q_t data for the N- and C-terminal halves of SsIGPS of the 3 folding pathways were compared with the data of all trajectories, as shown in *SI Appendix*, Fig. S6. The Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85 regions, representing the early and final folding stages, respectively, differ in their sources of frustration. The great similarity of the Q_i vs. Q_t plots at Q_t = 0.50 to 0.65 between all trajectories and the I_1 pathway (*SI Appendix*, Fig. S6 *A* and *B*) indicates that the global frustration at Q_t = 0.50 to 0.65 is primarily contributed by the frustration in the major I_1 channel.

However, the global frustration at Q_t = 0.75 to 0.85 is different, since no significant backtracking events were observed in the same region of all 3 folding channels (*SI Appendix*, Fig. S6 *B–D*). The 3 folding channels, differing in their assembly order of the protein, have very different values of $Q_{C\text{-ter}}$ and $Q_{N\text{-ter}}$ at different times. In the I_1 and I_2 channels, the C-terminal $(\alpha\beta)_{7-8}$ and the N-terminal $\alpha_0\beta_1$, respectively, are the last to fold. Before reaching the final stage of folding at Q_t = 0.75 to 0.85, the $Q_{C\text{-ter}}$ of the I_1 channel reaches ∼0.9, which is significantly larger than that of the I_2 channel (∼0.7). During most of the 2,000 time units the I_1 pathway trajectories (74 of 81) were trapped in the extremely stable I_1 state ($Q_t \sim 0.788$), which precludes sufficient sampling at $Q_t > 0.788$ in the I_1 channel and makes the I_2 and I_3 channels dominant in this region. Therefore, the backtracking of $Q_{C\text{-ter}}$ at Q_t = 0.75 to 0.85 is a result of the significantly smaller $Q_{C\text{-ter}}$ of the I_2 pathway and the sparse sampling of the I_1 pathway in this region.

Discussion

A combined experimental and computational study of the folding reaction of the SsIGPS TIM barrel has revealed insights into the structures of partially folded states and the potential role of frustration that occurs in simulations of the folding reaction.

Mechanistic Analysis. Previous experimental studies of the folding kinetics generated a 5-species model, $I_{BP} \rightleftharpoons U \rightleftharpoons I_A \rightleftharpoons I_B \rightleftharpoons N$ (Fig. 6) (11). Current native-centric simulations predict a more complex model involving partitioning between 3 different pathways to reach the native conformation (Fig. 5). The data from both models can be used to generate "kinetic species" plots to

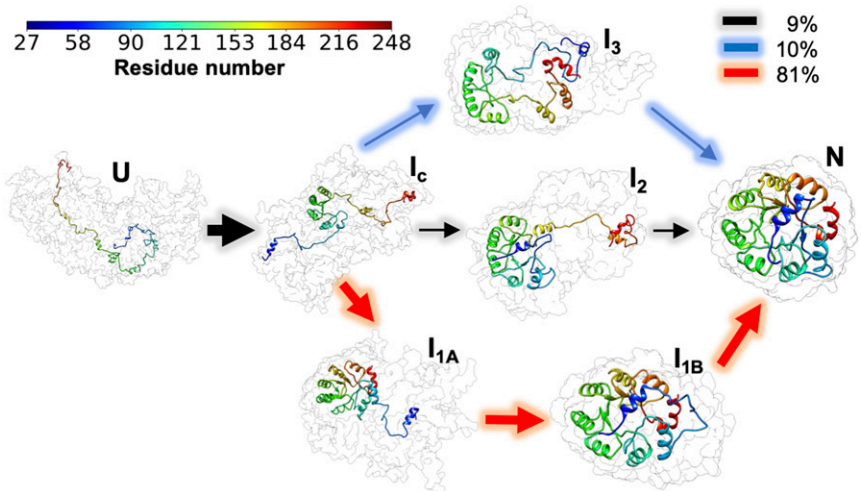

Fig. 5. Multiple folding pathways discovered by simulations, the upper right legend shows the transition probabilities from I_c to I_{1A}, I_2, and I_3. Additional structural details for I_{1A}, I_{1B}, I_2, and I_3 are shown in *SI Appendix*, Fig. S5. The gray contours show the overlay of ∼50 protein conformations, sampled from the corresponding states. See Movies S1–S3 for animations.

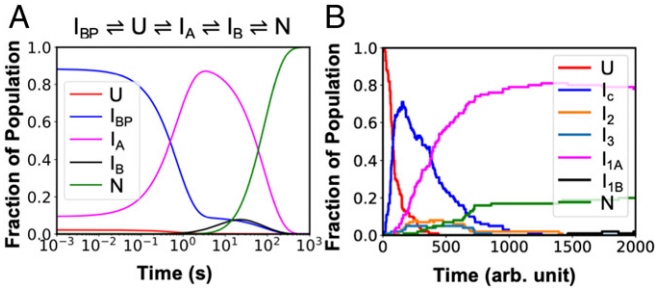

Fig. 6. Experimental (*A*) and simulated (*B*) populations of states vs. time.

compare the flow of material from the U state to the N state during a folding reaction (Fig. 6).

The 2 kinetic species plots are remarkably similar in several respects, but obviously differ with respect to the number of folding channels. The CF-FL and CF-SAXS measurements report the collapse of the unfolded chains within 50 μs; however, the distance distribution of the α_1-α_8 FRET pair (Fig. 3*A*) indicates the presence of 2 states. One state is more compact than the native state, implying a nonnative structure, and the other more expanded than native but more compact than the unfolded state. Because the experimental kinetic species plot (Fig. 6*A*) predicts the simultaneous presence of the I_{BP} and I_A states after a few milliseconds, with the I_{BP} state predominant, we presume that the overly compact state corresponds to I_A and the expanded state to I_{BP}. Examination of the predictions of the simulations after ∼200 time units (Fig. 6*B* and *SI Appendix*, Fig. S11) supports this interpretation when comparing the populations of the I_c and I_{1A} states. The subsequent increase in the population of the I_A state in experiments is also mimicked by the I_{1A} state in the simulations. Particularly striking is the correspondence of the very long lifetimes of both the I_A and I_{1A} states, consistent with their rate-limiting roles in folding by both experiment and simulations. Both experiments and the major refolding channel in the simulations then reveal a final intermediate, I_B and I_{1B}, respectively, before proceeding to the native state. The partitioning of I_c into 3 channels in the simulations is not evident in the experimental data. However, channels 2 and 3 each carry only ∼10% of the population and would be difficult to detect experimentally. Both experiments and the major refolding channel in the simulations then reveal a final intermediate, I_B and I_{1B}, respectively, before proceeding to the N state.

Structural Analysis. The pairwise and global dimensional analysis provided by CF-trFRET and CF-SAXS enables a direct comparison with the results of the simulations on the structures of the unfolded state, U, and the I_{BP}/I_c intermediates that appear in microseconds.

U state. SAXS measurements of the unfolded state in high concentrations of urea (Fig. 2*A*), when extrapolated to the absence of denaturant, yield an estimated Rg in water, 46 ± 5 Å, that is consistent with a random-coil ensemble for a chain of 226 amino acids (23). Remarkably, native-centric simulations of the U state (Fig. 4*B*) obtained the same estimate of Rg, ∼45 Å, and both approaches revealed the breadth of the unfolded manifold of conformers (Figs. 2*D* and 4*B*).

I_{BP}/I_c state. The trFRET data for the α_1-α_8 pair show that the microsecond folding reaction partitions into 2 distinct distributions with different degrees of contraction (Fig. 3*A*). One ensemble is more compact than that for the N state and the other much more expanded. Unfortunately, the limitations of FRET measurements of distance outside efficiencies of 0.2 to 0.8 preclude estimates of the relative populations of these distributions. As described above, these results are consistent with the previous global analysis of the folding of SsIGPS that found U partitioning into the I_{BP} and I_A states (Fig. 6*A*). Although the α_1-α_8 pair distances of the I_c state from simulations (*SI Appendix*, Fig. S9*A*) do not capture the overly compact conformations that appear after 50 μs (Fig. 3*A*), native centric simulations are incapable of detecting nonnative structures. As the time steps increase, more compact states appear at ∼20 and ∼10 Å (*SI Appendix*, Figs. S8*A* and S10*B*), reflecting the progression of the folding reaction toward the I_{1A}, I_{1B}, and N states. The α_3-α_4 FRET pair show a single distribution centered near that for the N state (Fig. 3*B*), but the greater breadth of which indicates a larger, dynamic ensemble. The SAXS data reveal a denaturant-independent Rg of 26 Å below 2 M urea after 150 μs (Fig. 2*A*), demonstrating that this species is not a collapsed form of the unfolded state (15). The Kratky plot after 150 μs (Fig. 2*C*) is not consistent with a fully globular structure, and the associated P(r) (Fig. 2*D*) shows a peak at ∼30 Å and a tail at longer distances.

The simulations show a remarkable degree of correspondence with experimental results for the structures that appear at the initial stage of folding for SsIGPS. The I_c intermediate has a native-like structure in the $(\beta\alpha)_{2-5}$ segment, consistent with the formation of a secondary structure detected by previous HDX-mass spectrometry (HDX-MS) experiments (8, 11) and the distance measured for the α_3-α_4 FRET pair. The N- and C-terminal segments are unstructured and give rise to the tail at high values of the P(r) (Fig. 2*D*), very similar to that seen in the SAXS data.

I_A/(I_{1A} and I_2). Although the present experimental study does not address the I_A intermediate, a previous HDX-MS study (8) found strong protection against exchange in the $(\beta\alpha)_{2-6}\beta_7$ region and a lack of protection in β_1 and β_8. As described above, the simulations show partitioning after the formation of I_c (Fig. 5). The dominant I_1 pathway involves the formation of a 7-stranded β-barrel, $\alpha_1(\beta\alpha)_{2-8}$, by locking out β_1 and α_0. The minor I_2 pathway formed a 6-stranded β-barrel by excluding $(\beta\alpha)_{7-8}$. In both cases, the nascent barrels appear to be stabilized by native-like interactions between α_1 and α_8 (Movies S1–S3). The exclusion of the N- and C-terminal β-strands would expose them to solvent and explain the lack of protection against HDX exchange.

I_B. The final intermediate in the experimental folding mechanism, I_B, has a fully formed β-barrel, providing protection against HDX in all 8 β-strands (8). This species appears after the rate-limiting step in folding and is only transiently populated as a minor species before the appearance of the N state (Fig. 6*A*). The simulations differ in that the I_{1B} state in channel 3 excludes β_1.

Frustration in Folding. Two major regions of topological frustration were found in the ensemble averaged analysis of the simulation, shown in the Q_i vs. Q_t plots (Fig. 4*C*). Multiple asynchronous folding pathways complicate the descriptions of the frustration in folding.

Frustration at Q_t = 0.50 to 0.65. The frustration at Q_t = 0.50 to 0.65, where the I_{BP}/I_c state persists, is mainly contributed by the backtracking events in the dominate I_1 channel. The backtracking event of the $Q_{N\text{-}ter}$ at Q_t = 0.50 (Fig. 4*C*) corresponds primarily to the unfolding of the $(\beta\alpha)_{1-2}$ element (Fig. 4*D*). This conclusion is consistent with experimental results in which some premature structures in the I_{BP}/I_c state are required to unfold before reaching the productive folding pathway. The 2 minor pathways I_2 and I_3 show different outcomes of the I_{BP}/I_c state in this region. The I_2 pathway shows backtracking in the C terminus while the I_3 pathway shows no obvious frustration. Interestingly, the different sources of frustration in the I_1 and I_2 pathways reflect alternative forms of an incomplete TIM barrel. The major I_1 pathway excludes the N terminus while the minor I_2 pathway excludes the C terminus. Both contain the central $(\beta\alpha)_{2-6}$ region that is protected against HDX (11) and, evidently, is capable of propagating structure in either direction.

Frustration at Q_t = 0.75 to 0.85. The frustration at Q_t = 0.75 to 0.85 is a combined result of the 3 folding pathways that differ in their assembly order of the protein and therefore does not represent

actual loss of structures. The global backtracking of $Q_{C\text{-}ter}$ at Q_t = 0.75 to 0.85 is caused mainly by the I_2 channel in which the C-terminal $(\alpha\beta)_{7-8}$ elements of the protein are the last to fold, thereby lowering the global $Q_{C\text{-}ter}$ value. Although no evidence of backtracking of the I_{1A} state ($Q_t \sim 0.79$) was found, it remains possible that I_{1A} may first partly unfold so that $\alpha_0\beta_1$ folds before the barrel closure to reduce the tremendous entropic cost required to make the transition from I_{1A} to I_{1B}.

Conclusions

A combined experimental and computational study of the folding reaction for a TIM barrel protein has yielded remarkable agreement between these studies' complementary views of a complex process. The mechanism defined by the major refolding pathway in simulations agrees closely with the mechanism determined by a variety of experiments. Striking similarities include the formation of stable structure in the central region of the sequence early in folding and a rate-limiting step before the formation of an 8-stranded β-barrel. Global and pairwise distance measurements of the early intermediate find a very similar degree of compactness, likely with disordered tails at both termini. In contrast to the experiments, the simulations reveal the presence of 2 minor channels that delineate alternative pathways to the native conformation. The frustration in folding detected by the simulations results in the formation of a pair of nascent TIM barrels that differ in the exclusion of either the N- or C-terminal segments of the protein, consistent with the presence of intermediates observed in experiments. Although it is likely that these incomplete barrels contain nonnative structures inaccessible to native centric simulations, the remarkable similarities in the minima on the experimental and computational folding free energy surfaces argue that they are dominated by native-like structures.

Materials and Methods

Protein Production. Protein was expressed in DE3 cells and purified using metal affinity, ion exchange, and sizing chromatography before labeling with 1,5-I-AEDANS. See *SI Appendix* for full details.

Fluorescence. Details of the time-correlated single-photon counting apparatus equipped with a microsecond continuous-flow mixer (Translume) have been described previously (18) and are discussed more completely in *SI Appendix*.

MEM. The 2D-MEM is used to obtain a distance distribution without any assumptions as to the number of lifetime components in the donor-only excited state decay, the shape of the distance distribution, or the number of subpopulations. Details of the mathematical framework used in this analysis are given in *SI Appendix*.

SAXS. SAXS measurements were performed at the BioCAT beamline at the Advanced Photon Source, Argonne, IL. Equilibrium SAXS measurements were performed by interfacing an autosampler running custom software to the standard quartz sample capillary (24). Kinetic experiments were performed as previously described (24) with the exception that flow to the quartz mixer for the kinetic experiments was controlled by syringe pumps (Harvard Apparatus) at a total flow rate of 4 to 5 $mL \cdot min^{-1}$. Scattering images were reduced using scripts provided by BioCAT and analyzed as previously described (18, 24).

Gō Model Simulations. The Cα-based native centric coarse-grained model of SsIGPS was generated by an in-house script using Protein Data Bank (PDB) code 2C3Z. Molecular dynamics simulations were performed using the CHARMM package (25). Snapshots were saved every 10,000 time steps (1 time unit). The 2,000 time units productive simulation sampled 100 independent trajectories at 230 K with unfolded initial configurations sampled at 510 K. See *SI Appendix* for further details.

ACKNOWLEDGMENTS. We thank Sagar Kathuria for expert assistance with the CF-SAXS and CF-FRET experiments. We thank Noah Cohen and Sagar Kathuria for helpful discussions and Divya Kapoor for help in creating the mutant constructs. This work was supported by National Institute of General Medical Sciences (NIGMS) Grant 5 R01 GM023303; the National Science Foundation (NSF) Division of Biological Infrastructure Grant 1353942; and the NSF Chemistry Division Grant 1506273 (to C.L.B.). This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. This project was supported by Grant 9 P41 GM103622 from the NIGMS of the National Institutes of Health (NIH). Use of the Pilatus 3 1M detector was provided by Grant 1S10OD018090-01 from NIGMS. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NIGMS or the NIH.

1. J. N. Onuchic, P. G. Wolynes, Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75 (2004).
2. R. D. Hills, Jr, C. L. Brooks III, Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *J. Mol. Biol.* **382**, 485–495 (2008).
3. A. I. Bartlett, S. E. Radford, An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat. Struct. Mol. Biol.* **16**, 582–588 (2009).
4. S. V. Kathuria, I. J. Day, L. A. Wallace, C. R. Matthews, Kinetic traps in the folding of β α-repeat proteins: CheY initially misfolds before accessing the native conformation. *J. Mol. Biol.* **382**, 467–484 (2008).
5. R. L. Baldwin, On-pathway versus off-pathway folding intermediates. *Fold. Des.* **1**, R1–R8 (1996).
6. C. Nishimura, H. J. Dyson, P. E. Wright, Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin. *J. Mol. Biol.* **355**, 139–156 (2006).
7. Y. Matsumura *et al.*, Transient helical structure during PI3K and Fyn SH3 domain folding. *J. Phys. Chem. B* **117**, 4836–4843 (2013).
8. Z. Gu, M. K. Rao, W. R. Forsyth, J. M. Finke, C. R. Matthews, Structural analysis of kinetic folding intermediates for a TIM barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and Gō model simulation. *J. Mol. Biol.* **374**, 528–546 (2007).
9. B. N. Gangadhara, J. M. Laine, S. V. Kathuria, F. Massi, C. R. Matthews, Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. *J. Mol. Biol.* **425**, 1065–1081 (2013).
10. Y. Wu, R. Vadrevu, S. Kathuria, X. Yang, C. R. Matthews, A tightly packed hydrophobic cluster directs the formation of an off-pathway sub-millisecond folding intermediate in the α subunit of tryptophan synthase, a TIM barrel protein. *J. Mol. Biol.* **366**, 1624–1638 (2007).
11. Z. Gu, J. A. Zitzewitz, C. R. Matthews, Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry: The roles of motif and sequence for the indole-3-glycerol phosphate synthase from Sulfolobus solfataricus. *J. Mol. Biol.* **368**, 582–594 (2007).
12. S. V. Kathuria, Y. H. Chan, R. P. Nobrega, A. Özen, C. R. Matthews, Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Sci.* **25**, 662–675 (2016).
13. S. V. Kathuria *et al.*, Advances in turbulent mixing techniques to study microsecond protein folding reactions. *Biopolymers* **99**, 888–896 (2013).
14. W. R. Forsyth, C. R. Matthews, Folding mechanism of indole-3-glycerol phosphate synthase from Sulfolobus solfataricus: A test of the conservation of folding mechanisms hypothesis in $(\beta\alpha)_8$ barrels. *J. Mol. Biol.* **320**, 1119–1133 (2002).
15. A. Borgia *et al.*, Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
16. R. P. Rambo, J. A. Tainer, Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **95**, 559–571 (2011).
17. A. T. N. Kumar, L. Zhu, J. F. Christian, A. A. Demidov, P. M. Champion, On the rate distribution analysis of kinetic data using the maximum entropy method: Applications to myoglobin relaxation on the nanosecond and femtosecond timescales. *J. Phys. Chem. B* **105**, 7847–7856 (2001).
18. Y. Wu, E. Kondrashkina, C. Kayatekin, C. R. Matthews, O. Bilsel, Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13367–13372 (2008).
19. W. R. Forsyth, O. Bilsel, Z. Gu, C. R. Matthews, Topology and sequence in the folding of a TIM barrel protein: Global analysis highlights partitioning between transient off-pathway and stable on-pathway folding intermediates in the complex folding mechanism of a $(\beta\alpha)_8$ barrel of unknown function from B. subtilis. *J. Mol. Biol.* **372**, 236–253 (2007).
20. J. Karanicolas, C. L. Brooks III, The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11**, 2351–2361 (2002).
21. J. Karanicolas, C. L. Brooks III, The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design? *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3954–3959 (2003).
22. X. Yang, S. V. Kathuria, R. Vadrevu, C. R. Matthews, βα-hairpin clamps brace βαβ modules and can make substantive contributions to the stability of TIM barrel proteins. *PLoS One* **4**, e7179 (2009).
23. J. E. Kohn *et al.*, Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12491–12496 (2004).
24. S. V. Kathuria *et al.*, Microsecond barrier-limited chain collapse observed by time-resolved FRET and SAXS. *J. Mol. Biol.* **426**, 1980–1994 (2014).
25. B. R. Brooks *et al.*, CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).