Modulation of frustration in folding by sequence permutation

Robert P. Nobrega
University of Massachusetts Medical School

Karunesh Arora
University of Michigan - Ann Arbor

Sagar V. Kathuria
University of Massachusetts Medical School

See next page for additional authors

Follow this and additional works at: http://escholarship.umassmed.edu/gsbs_sp

Part of the Biochemistry Commons, Molecular Biology Commons, and the Structural Biology Commons

Repository Citation
Nobrega, Robert P.; Arora, Karunesh; Kathuria, Sagar V.; Graceffa, Rita; Barrea, Raul A.; Guo, Liang; Chakravarthy, Srinivas; Bilsel, Osman; Irving, Thomas C.; Brooks, Charles L. 3rd; and Matthews, C. Robert, "Modulation of frustration in folding by sequence permutation" (2014). GSBS Student Publications. 1932.
http://escholarship.umassmed.edu/gsbs_sp/1932

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in GSBS Student Publications by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Modulation of frustration in folding by sequence permutation

Authors
Robert P. Nobrega, Karunesh Arora, Sagar V. Kathuria, Rita Graceffa, Raul A. Barrea, Liang Guo, Srinivas Chakravarthy, Osman Bilsel, Thomas C. Irving, Charles L. Brooks 3rd, and C. Robert Matthews

Comments
Publisher PDF posted as allowed by the publisher's author rights policy at http://www.pnas.org/site/aboutpnas/authorfaq.xhtml.

Rights and Permissions
Citation: Proc Natl Acad Sci U S A. 2014 Jul 22;111(29):10562-7. doi: 10.1073/pnas.1324230111. Epub 2014 Jul 7. Link to article on publisher's site

This article is available at eScholarship@UMMS: http://escholarship.umassmed.edu/gsbs_sp/1932
Modulation of frustration in folding by sequence permutation


*Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605; 2Department of Chemistry and Biophysics Program, University of Michigan, Ann Arbor, MI 48109; 3Institute for X-ray Physics, Georg-August-Universität Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany; and 4Biophysics Collaborative Access Team, Center for Synchrotron Radiation Research and Instrumentation and Department of Biological and Chemical Sciences, Illinois Institute of Technology, Chicago, IL 60616

Edited by David Baker, University of Washington, Seattle, WA, and approved June 13, 2014 (received for review December 30, 2013)

Folding of globular proteins can be envisioned as the contraction of a random coil unfolded state toward the native state on an energy surface rough with local minima trapping frustrated species. These substructures impede productive folding and can serve as nucleation sites for aggregation reactions. However, little is known about the relationship between frustration and its underlying sequence determinants. Chemotaxis response regulator Y (CheY), a 129-amino acid bacterial protein, has been shown previously to populate an off-pathway kinetic trap in the microsecond time range. The frustration has been ascribed to premature docking of the N- and C-terminal subdomains or, alternatively, to the formation of an unproductive local-in-sequence cluster of branched aliphatic side chains, isoleucine, leucine, and valine (ILV). The roles of the subdomains and ILV clusters in frustration were tested by altering the sequence connectivity using circular permutations. Surprisingly, the stability and buried surface area of the intermediate could be increased or decreased depending on the location of the terminus. Comparison with the results of small-angle X-ray-scattering experiments and simulations points to the accelerated formation of a more compact, on-pathway species for the more stable intermediate. The effect of chain connectivity in modulating the structures and stabilities of the early kinetic traps in CheY is better understood in terms of the ILV cluster model. However, the subdomain model captures the requirement for an intact N-terminal domain to access the native conformation. Chain entropy and aliphatic-rich sequences play crucial roles in biasing the early events leading to frustration in the folding of CheY.

Highly denatured states of globular proteins resemble statistical random coils when examined with low-resolution techniques such as X-ray scattering (1) and hydrodynamic analyses (2). However, a higher-resolution view provided by experimental models (3–6) and simulations (7) shows that the conformational ensemble is biased toward low-contact-order (CO) structures, e.g., \( \alpha \)-helices, \( \beta \)-turns, and \( \beta \)-hairpins, which form and melt in less than a few microseconds. During folding, these nascent structures presumably coalesce into higher-order assemblies of ever-increasing free energy until reaching the transition-state ensemble (TSE) that leads to the native conformation. From another perspective, this assembly process mediates a global collapse of the chain in an unfavorable solvent (8). Landscape theory (9) posits that, in the simplest scenario, native-like substructures appear and lead without pause to the TSE and the native conformation in an apparent two-state fashion. However, simulations have found that topological frustration, e.g., the premature formation of a substructure that impedes access to the productive TSE, can lead to the accumulation of intermediates (I) that must unfold to some extent to traverse the folding reaction coordinate successfully (8, 10, 11). Experimental and computational studies on the folding of the \( \alpha \)-subunit of Trp synthase (12, 13), the chemotaxis response regulator Y (CheY) (10, 14), a pair of apo-flavodoxins (8, 15, 16), and tandem titan domains (17) revealed frustration in the form of off-pathway intermediates (I_{OFF}). Thus, as-yet-unexplored aspects of sequence and structure can add complexity to folding reactions.

The observed inverse relationship between CO and folding rate constant (18) implies that elements of secondary structure that are near in sequence and near in space will associate preferentially over those that are distant in sequence. However, if such low-CO substructures are not involved in the productive TSE, they could serve as sources of frustration. A case in point is CheY, a member of the very common flavodoxin-fold family with its classic \( \alpha/\beta \)-sandwich architecture. The (\( \beta_1\alpha_5\)) motif displays the \( \alpha_1 \) and \( \alpha_5 \) helices on one face of the parallel \( \beta \)-sheet and the \( \alpha_2 \), \( \alpha_3 \), and \( \alpha_4 \) helices on the opposing face (Fig. 1 A and B). The proposed kinetic folding mechanism (Fig. 1C) (14) involves two parallel folding channels defined by the \( cis \) and \( trans \) isomers of the prolyl peptide bond between K109/P110. The unfolded proteins (U) in both the major \( trans \) (U_t (90%)) and minor \( cis \) (U_c (10%)) channels sample an off-pathway submillisecond intermediate (I_{OFF}), I_m, and I_P, respectively, before the rate-limiting isomerization reaction in the I_{OFF}\rightarrow I_m step. I_m unfolds to the U_c state before accessing the productive TSE leading to the native conformation (N) in the U_c\rightarrow N_c step. Further complicating the mechanism is an on-pathway intermediate, I_{ON}, between U_t and N_t that has been observed by equilibrium NMR.

Significance

Folding mechanisms of large proteins are often complicated by the existence of kinetic traps that impede progress toward the native conformation. We have tested the role of chain connectivity in creating such traps by permuting the sequence of a small \( \alpha/\beta/\alpha \)-sandwich protein, the chemotaxis response regulator Y. An approach combining experimental and native-centric simulations reveals that chain entropy and aliphatic-rich sequences conspire to create frustrated species whose structures and stabilities vary with connectivity. The initial events in folding reflect not a random collapse driven by the hydrophobic effect but rather the accumulation of substructures favored by low-contact-order nonpolar interactions in the polypeptide. The conserved global free-energy minimum of the native conformation ultimately resolves these early frustrations in folding.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1R.P.N. and K.A. contributed equally to this work.

2To 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.1324230111/-/DCSupplemental.
We introduced sequence permutations that can bias the initial stages of folding toward the productive TSE and away from kinetic traps.

Results

**Permutations Differentially Affect the Secondary Structures of the Folded States of the Permutants.** We introduced sequence permutations into the F14N variant of CheY, denoted “CheY*”, to increase the stability of the platform and its tolerance for the introduction of the linker and the new termini; the folding mechanism for CheY* is unchanged from the WT protein (20). The new N termini for Cpb2, Cpb3, and Cpb4 become D38, D64, and E89, respectively. An additional glycine residue at position –1 is a remnant of the cleaved hexahistidine affinity tag (SI Methods). The far-UV CD spectrum of Cpb2 is markedly different from CheY* (Fig. S1), with the relative intensities of the double minima at ~210 and ~222 nm reversed from those of the CheY* protein. Unfortunately, the substantial perturbation of the secondary structure precluded Gō-model simulations that rely on knowledge of the native structure. The CD spectra of Cpb3 and Cpb4 display the same relative double minima as CheY*; the spectrum of Cpb4 is coincident with CheY*, and Cpb3 is reduced in amplitude by ~15% (Fig. S1). Although the secondary structure of Cpb3 appears to fray to some extent, the basic βαβ architecture is preserved. Therefore, both Cpb3 and Cpb4 were deemed to be good candidates for a combined experimental and computational analysis of their folding mechanisms.

Stability Analysis of the Permutants.** The concerted disruption of secondary and tertiary structures with increasing concentrations of urea revealed an apparent two-state process, N=U, for CheY* and Cpb2 (Fig. 2 and Table S1). Fits of the data to a linear dependence of free energy of folding on the denaturant concentration (23) showed that the stabilities varied from 2.11 kcal·mol⁻¹ for Cpb2 to 8.0 kcal·mol⁻¹ for CheY* protein (Table S1). The denaturant dependence of the free energy of folding, the m-value [a measure of the change in buried surface area (24)], varied from 0.77 kcal·mol⁻¹·M urea⁻¹ for Cpb2 to 1.99 kcal·mol⁻¹·M urea⁻¹ for CheY*.

Nobrega et al.  |  PNAS  |  July 22, 2014  |  vol. 111  |  no. 29  |  10563

Fig. 1. (A) Topology diagram of CheY. The N-terminal folding subdomain is highlighted in yellow, and the C-terminal folding subdomain is highlighted in blue. The effects of each permutation on the continuity of cluster 1 (blue), and cluster 2 (red) are shown. (B) Clusters of ILV residues are superimposed on the crystal structure of CheY (Protein Data Bank ID code: 3CHY). Cluster 1 (blue) has a lower CO and resides on the α3β3α4 side of the central β-sheet. The larger cluster 2 (red) contains high-CO contacts and resides on the α1β3 side of the β-sheet. (C) The folding mechanism of WT CheY. The major pathway is highlighted in red. The U₁ → N₃, step, involving the on-pathway intermediate, is designated by the triple dots.

Measurements (19) and in Gō-model simulations (10) but not by circular dichroism (CD) or flow fluorescence (FL) experiments.

Mutational analysis (20, 21) has revealed a nucleation-condensation folding mechanism for CheY in which the N-terminal subdomain [residues 1–70, (α)α1–2β3] serves as the nucleus for the subsequent condensation of the C-terminal subdomain [residues 70–129, α3(α)4–5] (Fig. 1A). However, native-centric simulations identified contacts between the N- and C-terminal subdomains, centered around (α)3β3–4, early in folding that are incompatible with access to the productive TSE and that lead to frustration in the folding mechanism (10). Another perspective is provided by the branched chain aliphatic side chains (BASIC) hypothesis, which supposes that large clusters of isoleucine, leucine, and valine (ILV) side chains serve as cores of stability in folding intermediates (11, 14). Both these clusters have been shown to have a high contact density (22). CheY has two ILV clusters, each serving to fuse the surface helices to each other and to the central β-sheet; the larger cluster 2 contains 15 side chains and links the β-strands to α1 and α5. The sequence spanned by the smaller cluster agrees closely with the (α)3–4 segment identified as the source of frustration in the simulations and, importantly, involves only low-CO contacts. If cluster 1 were to form early and, by sequestering β3, impede the development of the productive TSE in the N-terminal subdomain, (α)1–2β3, the BASIC hypothesis would provide an alternative explanation for frustration in the folding of CheY.

Permutations in the sequence of CheY provide a means to compare the subdomain model and the ILV cluster models as explanations for the frustration in folding detected by simulations and experiments. By fusing the natural N and C termini with a short linker peptide (Gly-Ala-Gly) and inserting new termini in the loops after β2, β3, and β4, it is possible to cleave within the N-terminal subdomain (Cpb2), between the subdomains (Cpb3), and within the C-terminal subdomain (Cpb4). Related to the ILV clusters, Cpb2 cleaves cluster 1 and leaves cluster 2 largely intact, Cpb3 cleaves both clusters, and Cpb4 cleaves only cluster 2 (Fig. 1A). Our simulations and experiments on these permutants show that aspects of both models describe the relationships between sequence, structure, and frustration in the folding of CheY. The results also show that frustration can be modulated by sequence permutations that can bias the initial
a two-state model, kinetic analysis (see below) revealed the presence of a stable intermediate and dictated a three-state model. The melting temperatures estimated from the temperature dependence of heat capacities calculated by the simulations (Fig. S3) are in the same rank order as the midpoint temperatures in the Trp fluorescence (Fig. 2). Cpβ3 < CheY* < Cpβ4. Experimental thermal melts by both DSC and CD were irreversible, and a reliable experimental measurement of the melting temperatures could not be obtained. Further experiments on Cppβ2 were not pursued.

**Kinetic Analysis of Permutant Folding.** We monitored the dynamic responses of the permutants to rapid changes in the denaturant concentration in the microseconds-to-hundreds of seconds time range with a combination of continuous-flow (CF), stopped-flow (SF), and manual-mixing (MM) techniques interfaced with FL, circular dichroism (CD), and small-angle X-ray-scattering (SAXS) detection. For CheY*, a large-amplitude FL phase occurs within the 25-μs dead time of CF refolding, followed by a small-amplitude phase lasting several hundred microseconds. The subsequent formation of the native state occurs in hundreds of seconds and has been attributed to the trans → cis isomerization of the K109–P110 peptide bond (14) (Fig. S4). Unfortunately, refolding along the cis channel for the permutants could not be resolved because of its small amplitude in direct refolding experiments and interrupted unfolding experiments. A pair of unfolding reactions were observed in the seconds-to-hundreds of seconds time range; the interconversion of the native cisP110 conformer to its trans counterpart, Nn → Nf, controls unfolding in the transition zone and the direct unfolding of the native cisP110 to the unfolded cisP110, Nn → Un, controls unfolding at high denaturant concentrations. Similar overall responses were observed for Cppβ3 and Cpβ4, with the exception that the direct unfolding of the native cisP110 conformer was accelerated for Cpβ4.

**Stability and Secondary Structure of Submillisecond Intermediate States.** The orders of magnitude in time (from microseconds to hundreds of seconds) separating the folding reactions for all three proteins enabled us to measure the stability of the product of the microsecond reaction, In, and its CD spectrum. By plotting the ellipticity at 222 nm after 5 ms of refolding in varying final denaturant concentrations, the stability can be estimated by fitting the resulting titration curve to a two-state model (Fig. 2). The Inp species for Cppβ3 is significantly less stable than for CheY*, 0.84 kcal mol⁻¹ vs. 2.02 kcal mol⁻¹, and the m-value is also decreased (Table S1). Very surprisingly, Inp stability is much greater for Cpβ4 (4.31 kcal mol⁻¹) than for CheY*, and the m-value is increased (Table S1). Comparison of the denaturation curves for folded Cppβ4 and its Inp species (Fig. 2) shows that the two curves overlap between 5 and 5 M urea. By fixing the thermodynamic parameters for the Inp = U reaction to those extracted from the burst-phase titration data, the stability and m-value for the Na⁺=U reaction could be estimated by fitting the equilibrium titration data for Cpβ4 for a three-state model. The difference in free energy between its native and unfolded forms is 8.19 kcal mol⁻¹, and the m-value is 1.80 kcal mol⁻¹ M⁻¹, comparable to that of CheY*.

We obtained the CD spectra of the Inp species by refolding jumps to the same final urea concentration in the folded baseline and varying the detection wavelengths in the far-UV range. The Inp species for CheY*, Cppβ3, and Cpβ4 recover 85%, 80%, and 90% of their native ellipticities at 222 nm within 5 ms (Fig. 2). The subtle but significant differences previously observed between the Inp and native states of CheY* (14) (Fig. S5) indicate that the aromatic side chains have not yet attained their native packing. In contrast, the very similar shapes of the spectra for the Inp and native states of Cpβ4 and Cppβ3 show that an exciton coupling, likely between the side chains in a cluster of phenylalanines on the α1/α5 side of the β-sheet (Fig. S5 C and D), is present in the Inp state for both permutants.

**Compaction of CheY* and Cpβ4 by CF-SAXS.** The very surprising increases in the stability and the apparent compaction for the Inp species for Cpβ4, the latter implied by the increased m-value for its urea melt, motivated us to measure its radius of gyration (Rg) in the ~100-μs-to-1-ms time range by CF-SAXS. The urea-denatured states of CheY* and Cpβ4 display Rg's of ~35 Å, slightly smaller than predicted for space-filling random coils of 129 amino acids, 38 Å (25). CheY* collapses to an apparent Rg of ~25 Å within the ~100-μs dead time, experiences a further compaction to ~23 Å by 1 ms, and ultimately contracts to an Rg of 15 Å in the native conformation (Fig. 3A). In distinct contrast, Cpβ4 collapses to a near-native Rg ~18 Å within ~100 μs and remains unchanged after 2.4 ms before contracting to the 15.5 Å Rg of the native state (Fig. 3A). Although the change in connectivity does not have a discernible effect on the size of the unfolded ensemble, the cleavage of the chain after P4 and the fusion of the natural N and C termini cause Cpβ4 to collapse more rapidly to a near-native Rg.

**Topological Frustration by Simulations.** The significant differences in the stabilities of the Inp species of these proteins are surprising, given the similarity of the kinetic responses observed. Unfortunately, the small amplitude of the refolding reaction along the cis channel precluded the use of global analysis to resolve the folding mechanism of the permutant proteins. Therefore we used Gō-model simulations to resolve the underlying structural basis of the differences in the Inp stability and infer the kinetic model that is most consistent with the experimental observables. Previous experimental work has concluded that the Inp is not a consequence of the proline isomerization reaction (14, 21). Likewise, in computational work where the trans geometry was enforced via harmonic restraints, CheY* still was able to access the folded state from the unfolded configurations. Although, the folded state is destabilized by 2.1 kcal mol⁻¹ relative to flexible Pro110 (26), the relative energy landscapes of the cis and trans channels are similar in the native and intermediate states (10, 11).

Despite using a model in which native interactions are predominantly favored, the model can capture frustration arising from the formation of native interactions in an incorrect order (27). Fig. 4 shows the influence of chain connectivity on the topological frustration as deduced from folding simulations of CheY* and its circular permutants. Our results are consistent with those reported earlier (10) and show that the folding of CheY*
proceeds with significant frustration that arises from the competition of interactions between N-terminal, C-terminal, and interfacing native contacts. At the fraction of total native contacts formed (Q_{total} = 0.4), local unfolding or backtracking of interfacing contacts (negative slope) between the N and C termini coincides with the sudden increase in the contacts of the N-terminal subdomain (Fig. 4 A and D). These prematurely formed contacts in the C subdomain partially unfold before folding proceeds to the native conformation. Similar results are observed for Cpf4; however, the interfacing frustration is markedly reduced (Fig. 4 B and E). In Cpf3 this interfacing frustration is absent (Fig. 4 C and F) because the new terminus disrupts the frustrated region. These results are consistent with the ILV cluster model for folding, in that the WT connectivity is driven to fold to the off-pathway I_{BP} species by the premature formation of cluster 1 spanning the interfacing contacts (14). The local connectivity of the larger cluster 2 in Cpf4 enables it to outcompete the formation of cluster 1.

Notably, a minor restructuring event in the N-terminal subdomain is observed late in folding at Q_{total} = 0.6 in all connectivity. This second event corresponds to the loosening of structure that is observed initially in the folding of α-helices before final maturation of the tertiary structure and is not comparable to early frustration (27).

Kinetic Simulations. More detailed structural insights into the folding mechanisms are gleaned through simulations from the time evolution of R_{g} and the corresponding time courses of the mean fraction of secondary structure contacts formed for the representative folding trajectories of CheY* and permutants. Cpf4 collapses faster than CheY* (Fig. 3B) before both approach a common R_{g} of ~14 Å in their respective native conformations. Examination of individual trajectories for CheY* and Cpf4 (Fig. S6) revealed pauses, reflecting the transient occupancy of partially folded states with discrete R_{g} values. Of 100 kinetic trajectories, only about half pass through this intermediate and persist long enough to be observable. Therefore, the intermediate can be regarded as a nonobligate I_{BP}. Statistical analysis suggests that the intermediate for Cpf4 is slightly more compact (20.2 vs. 21.3 Å) and undergoing even distribution of time units, and disappears sooner (104 vs. 151 time units) than its CheY* counterpart (Fig. 3B and Table S2). The R_{g} for these intermediates are in remarkably good agreement with those observed by SAXS after 1 ms of folding, ~25 Å for CheY* and ~18 Å for Cpf4 (Fig. 3A). The differences in the folding kinetics of CheY* and the permutants may reflect the extent of frustration that arises during the folding of each system.

To characterize the intermediates structurally, we extracted structures sampled during kinetic folding simulations that fall within 20 Å < R_{g} < 22 Å and measured the probability of forming native contacts in this ensemble. The results for CheY* are consistent with previous work (10) in which the N_{heptad} was identified as the structured region encompassing (ββ)-1–3β4 (Fig. S7A). Further, a subsection of the N_{heptad} with the highest probability of contact formation is apparent at the subdomain interface, β3–β4, a region previously described as an area where topological frustration is present (5). Through a similar analysis of Cpf4 (Fig. S7B) and Cpfβ (Fig. S7C), the differences in chain connectivity were found to have structural repercussions on the early-folding intermediates. The probability of forming contacts in the frustrated region is diminished as the N_{heptad} is lengthened to include α5/β5.

Discussion

CheY is a member of the flavodoxin fold family of proteins whose α/β/a sandwich architecture represents one of the more common motifs in biology. Unlike the flavodoxins, CheY has a conserved cis proline that controls the access to the native conformation (14, 20). Like CheY, however, a pair of homologous flavodoxins sample a kinetic trap before successfully traversing the productive TSE (8, 16). Elucidating the molecular basis for the frustration in folding for CheY has implications for an entire motif.

CheY*: The results presented here on the F14N variant of CheY are consistent with previous experimental and computational work on the WT protein (10, 14). By Go-model simulations, topological frustration arises at the subdomain interface before partially unfolding to resume folding from the N→C terminus. This result is consistent with experimental data that show nonnative Phe packing in the I_{BP} species (Fig. S5). The nonnative packing of I_{BP} along with the backtracking observed by simulations (Fig. 4 A and D; see also ref. 14) and the negative m-value observed through global analysis of experimental data (14) argue that CheY populates an off-pathway kinetic trap, I_{OFF}. Mechanistic details gleaned from the simulations suggest that low-CO ILV contacts in cluster 1 drive early folding events and lead to the premature formation of the subdomain interface.

Cpfβ2: By introducing new termini between β2 and α2, the Cpfβ2 permutant cleaves the N-terminal subdomain while leaving cluster 1 essentially intact and cluster 2 discontinuous. The observation that Cpfβ2 is incapable of adopting the CheY* fold demonstrates a key role for the intact N-terminal subdomain because cluster 2 is discontinuous in CheY* and in all three permutants. This conclusion is consistent with the results of a previous mutational analysis, which found the N-terminal subdomain to be a central feature of the productive TSE (21).

Cpfβ3: The introduction of new termini between β3 and α3 leaves the two subdomains intact but cleaves both ILV clusters and the N_{heptad}. Notably, the FL and CD titrations are noncoincident, suggesting that multiple species are present before
the global unfolding reaction. However, because the kinetic response is similar to that of CheY* under strongly unfolding conditions, Cpβ3 transverses the same barriers as CheY* (Fig. S4). The additional faster phase in unfolding may reflect a small fraction of the protein moving through a parallel channel in a limited range of unfolding conditions.

Although the amplitude of the CD spectrum of the Ibp species for Cpβ3 is decreased by only \(\sim 15\%\) from its CheY* counterpart, the stability is reduced markedly, from 2.02 kcal mol\(^{-1}\) for Ibp\(_{in}\) in CheY* to 0.84 kcal mol\(^{-1}\) in Cpβ3 (Table S1), and the m-value is reduced from 0.83 to 0.59 kcal mol\(^{-1}\) M\(^{-1}\). We attribute the decreased stability of Ibp\(_{in}\) to the cleavage of cluster 1, postulated to be a key stabilizing component of the Ibp\(_{in}\) species for WT CheY* (14). Interestingly, the loss in stability is accompanied by native-like packing of the Phe cluster on the \(\alpha\)15 face of the \(\beta\)-sheet (Fig. 2D).

Simulations show the elimination of the interfacial frustration of the subdomains for Cpβ3, which would be expected if \(\beta3\) and \(\beta4\) are segregated to opposite ends of the chain. The absence of early frustration in the Cpβ3 simulations may reflect the marginal stability of the Ibp species, as has been observed previously for a CheY homolog, NT-NtrC (14). In contrast to CheY*, frustration in Cpβ3 arises late in folding around the \(\beta1/\beta3/\beta5\) interface on the opposite face of the \(\beta\)-sheet (Fig. S8). The high Q values where this frustration occurs are not consistent with the small m-value for the Ibp\(_{in}\) species for Cpβ3 and likely reflect annealing reactions often seen in the late stages of folding in Go-model simulations when helix repacking often occurs.

The structural basis for the altered folding properties in Cpβ3 can also be visualized in 2D contact probability maps derived from the simulations (Fig. S7C). For its Ioff\(_{BP}\) species, CheY* has a high probability of contacts in the \(\alpha2/\beta1/\beta3/\beta4\) region, but Cpβ3 does not. Indeed, the region of high probability of native contacts in Cpβ3 shifts to the \(\beta\)1 and \(\beta5\)5 segments that are covalently linked by permutation of the sequence.

**Cpβ4:** The introduction of new termini between \(\beta4\) and \(\alpha4\) in Cpβ4 cleaves the C-terminal subdomain while leaving cluster 1 intact and cluster 2 discontinuous. The coincidence of the far-UV CD spectra of CheY* and Cpβ4 (Fig. S1) shows that an intact C-terminal subdomain is not essential for proper folding and is in agreement with the view that the C-terminal subdomain forms after the TSE (21). The resultant Ibp\(_{in}\) species folds more rapidly, is both more stable and more compact than CheY*, and has native-like packing of its phenylalanine cluster. The increased stability of Ibp\(_{in}\) provides a logical explanation for the accelerated unfolding reaction, via the Hammond effect (Fig. 5) and argues for its assignment as an on-pathway intermediate. These surprising experimental results are in very good agreement with the predictions of decreased frustration from an off-pathway intermediate and a more compact on-pathway intermediate including \(\beta1, \alpha1, \beta5, \alpha5\) in the simulations.

The 2D contact probability map of the Cpβ4 folding intermediate reveals an intact Nheptad and a high probability for contacts between the covalently connected \(\beta1/\beta5\)5 sequences. The linkage of the natural termini leads to the preferential formation and stabilization of a species that corresponds to the Ioff\(_{BP}\) for CheY*. The decreased frustration for Cpβ4 likely reflects both the destabilization of the C-terminal subdomain via cleavage and the increased competition from the more rapidly forming and stable extended Nheptad, including the \(\beta1/\beta5\)5 complex.

**Early Folding Events by CF-SAXS, Simulations, and CF-FL.** The faster collapse of unfolded Cpβ4 observed by CF-SAXS (Fig. 3A) and simulations (Fig. 3B) is not reflected in the CF-FL data, which found essentially identical relaxation times for Cpβ4 and CheY* (Fig. S4). The discrepancy can be traced to the small m-value for the 300-μs phase and the implied small change in buried surface area accompanying this reaction. The commonality of the relaxation time of this phase for Cpβ3, Cpβ4, and CheY* strongly suggests a local folding event at the single Trp residue that does not reflect the global collapse monitored by CF-SAXS and simulations.

**Modulation of the Folding Landscape by Permutations.** Both experiments on and simulations of CheY*, Cpβ3, and Cpβ4 reveal that the initial events in the folding are dictated by the connectivity of the chain. In another case, Cpβ2, altering the chain connectivity leads to a distinctly different but well-defined thermodynamic state. The combined results for those sequences that can attain the wild-type native conformation can be displayed on a reaction coordinate diagram shown in Fig. 5D; the proposed structured elements for the various species are shown in Fig. 5A–C.

The path from the unfolded state to the respective intermediates for CheY*, Cpβ3, and Cpβ4 is controlled by preferred interactions between low-CO elements of secondary structure. The varying structures, stabilities, and buried surface areas for these partially folded states can be understood in terms of the thermodynamic compulsion to minimize the chain entropy penalty and maximize the participation of their resident aliphatic side chains in one of two ILV clusters located on either face of the central \(\beta\)-sheet. For CheY*, cluster 1 forms early and stabilizes Ioff\(_{BP}\). For Cpβ3, cluster 1 is cleaved, and a fraction of cluster 2 drives the formation of a poorly folded fragile Ioff\(_{BP}\). For Cpβ4, the C-terminal elements of cluster 2 reinforce the \(\alpha\)C in a limited range of unfolding conditions.

**Subdomain vs. ILV Cluster Model for the Folding of CheY.** The totality of the results suggests that the ILV cluster model provides the
more parsimonious and complete description of the early events in folding but that the subdomain model better captures the crucial TSE required to access the proper native fold. In other words, low-CO clusters of ILV residues can strongly influence the early stages of folding before subdomain and global cooperativity engage expanding portions of the sequence to reach the native conformation.

**Perspective.** Chain entropy plays a crucial role in defining the energies and structures of partially folded states on the folding free-energy surface of CheY. Thus, frustration can be modulated and productive folding favored by altering the sequence connectivity and thereby, the local chain entropy. The local-in-space topology of the repeat proteins, including the Rossmann-fold, triosephosphate isomerase barrels, and the flavodoxin/CheY folds, make them prime candidates for frustration in the early stages of folding. The associated partially folded states not only may impede the folding reaction but also may serve to nucleate aggregation reactions in pathological sequence variants. Recognition of the early events in folding and the partially folded structures that they produce provides a rational basis for the design of small molecules that might inhibit aggregation by binding at the interfaces of these nascent kernels of structure.

**Methods**

**Thermodynamic and Kinetic Experiments and Analysis.** Details regarding protein expression, purification, and thermodynamic and kinetic characterization have been described previously (14). For details see *SI Methods*.

**Equilibrium SAXS.** Equilibrium measurements were collected as previously described (28). The protein concentration was 1.5 mg mL⁻¹ in 10 mM potassium phosphate buffer at pH 7.0 and 25 °C.

**CF-SAXS.** CF-SAXS measurements were made as previously described (29). The total flow rate was 20 mL min⁻¹ using a 1:10 dilution of the unfolded protein for a final protein concentration of 1.5 mg mL⁻¹ in 10 mM potassium phosphate, 8 M urea at pH 7.0 and refolding with 0 M urea buffer.

**Go-Model Simulations.** System preparation and model. Cpf1 and Cpf3 were modeled on the crystal structure of WT CheY from *Escherichia coli* (Protein Data Bank ID code: 3CHY) (30). Models of both permutants were constructed by joining together the N and C termini with a Gly-Ala-Gly peptide and cleaving the bond between residues 63 and 64 and between residues 88 and 89 for the Cpf3 and Cpf4 permutants, respectively. The protein-folding simulations were performed with an unrestrained prolyl-bond geometry using a coarse-grained model developed by Karanicolas and Brooks (31). See *SI Methods* for further details.

**Molecular dynamics protocol.** Molecular dynamics simulations were performed using the CHARMM macromolecular mechanics package (32). All models were evolved through Langevin dynamics, by using a friction coefficient of 1.36 ps⁻¹ and a molecular dynamics time step of 22 fs. The virtual bond lengths were kept fixed using the SHAKE algorithm. For each permutant, 100 independent folding/melting runs were performed for 2 × 10⁶ dynamics steps at 0.87 T, where T is the folding transition temperature estimated as a temperature corresponding to the peak in the specific heat curve, C(T) (Fig. S3). See *SI Methods* for details.

**ACKNOWLEDGMENTS.** We thank Jill Zitzewitz and Noah Cohen for helpful discussions, Omella Biscigia for helping with protein preparation, and Ronald Hills for help in revising the paper. This work was supported by the National Institutes of Health (NIH) through the Center for Multi-Scale Modeling Tools for Structural Biology Grant RR012255, the National Science Foundation through the Center for Theoretical Biological Physics Grant Grant No. DMR-0555561, the Division of Molecular and Cellular Biosciences Grant MCB1121942, National Center for Research Resources Grant 2P41RR08630-17, and NIH/National Institute of General Medical Sciences Grant P01 GM103622-17. Use of the Advanced Photon Source, an Office of Science User Facility operated for the US Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the US DOE under Contract No. DE-AC02-06CH11357. This project was supported by Grant 9 P41 GM103622 from the National Institute of General Medical Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Institute of General Medical Sciences or the National Institutes of Health.