In Silico Protein Fragmentation Reveals the Importance of Critical Nuclei on Domain Reassembly

Lydia M. Contreras Martínez, Ernesto E. Borrero Quintana, Fernando A. Escobedo, and Matthew P. DeLisa
School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York

ABSTRACT Protein complementation assays (PCAs) based on split protein fragments have become powerful tools that facilitate the study and engineering of intracellular protein-protein interactions. These assays are based on the observation that a given protein can be split into two inactive fragments and these fragments can reassemble into the original properly folded and functional structure. However, one experimentally observed limitation of PCA systems is that the folding of a protein from its fragments is dramatically slower relative to that of the unsplit parent protein. This is due in part to a poor understanding of how PCA design parameters such as split site position in the primary sequence and size of the resulting fragments contribute to the efficiency of protein reassembly. We used a minimalist on-lattice model to analyze how the dynamics of the reassembly process for two model proteins was affected by the location of the split site. Our results demonstrate that the balanced distribution of the "folding nucleus," a subset of residues that are critical to the formation of the transition state leading to productive folding, between protein fragments is key to their reassembly.

INTRODUCTION

Recent advances in molecular biology techniques have led to the development of many powerful research tools that have been key in providing detailed knowledge of the principles underlying highly specific interactions between cellular proteins. Of particular note is the protein fragment complementation assay (PCA), wherein a reporter protein is split into individual fragments that by themselves remain inactive but upon reassembly under the appropriate cellular conditions yield the original, properly folded and active protein structure. For example, the yeast two-hydrid system, based on the functional reconstitution of the split Gal-4 transcriptional activator (1), has facilitated the systematic determination of proteome-scale protein-protein interaction networks within numerous organisms, including humans (2), *Drosophila melanogaster* (3), *Caenorhabditis elegans* (4), *Saccharomyces cerevisiae* (5,6), vaccinia virus (7), and *Escherichia coli* bacteriophage T7 (8).

The increasing interest in protein-protein interactions has motivated the search for additional split reporter proteins that can be used for different applications and in other systems besides yeast (9). Examples include split green fluorescent protein (GFP) and its spectral variants yellow FP and cyan FP (10,11), ubiquitin (12), murine dihydrofolate reductase (DHFR) (13), β-lactamase (14,15), and firefly luciferase (16). The use of these split proteins is highly convenient, since the reconstituted activity of each is directly measurable by fluorescence or other well-established enzymatic assay. Numerous successes notwithstanding (17,18), the use of split proteins can be limited in usefulness because of the slow folding kinetics and formation of misfolded aggregates associated with the reassembly process of the fragments (11,17). For instance, whereas GFP activity can be detected in minutes, the two split fragments that result when the protein is dissected near the middle of the sequence fail to associate and reassemble when expressed in bacteria (11). A similar drawback has also been observed in other split systems like DHFR, β-lactamase, and ubiquitin, where folding is dramatically (or completely) inhibited upon protein fragmentation. In most cases, the addition of two interacting proteins to the split halves dramatically improves the kinetics of split protein reassembly, presumably by nucleating the reassembly reaction (11). However, even when fragments are each fused to strongly interacting leucine zippers ($K_D \approx$ 1–20 μM), folding and activity of the reconstituted protein are achieved only after 1–2 days (19). This inefficiency hinders the effective application of these detection systems on biologically relevant timescales. In an effort to increase the self-assembly efficiency of protein fragments in the absence of any interacting partners, a number of strategies have been employed, including 1), the identification of ''permissive'' split sites along the protein sequence using circular permutation (20,21), structure-guided design (14,22), or bioinformatic and theoretical analyses (23); and 2), the optimization of a target sequence for more efficient splitting/reassembly using directed evolution (24,25). In the majority of cases, split sites are often selected in regions away from the catalytic site, in areas containing flexible loops that can typically tolerate amino acid insertions, or in linker regions that separate naturally occurring functional domains (17).

Submitted August 11, 2007, and accepted for publication October 9, 2007.

Lydia M. Contreras Martínez and Ernesto E. Borrero Quintana contributed equally to this work.

Address reprint requests to Fernando A. Escobedo, School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853. Tel.: 607-255-8243; Fax: 607-255-9166; E-mail: fe13@cornell.edu; or to Matthew P. DeLisa at the same address. Tel.: 607-254-8560; Fax: 607-255-9166; E-mail: md255@cornell.edu.

Editor: Ron Elber.

0006-3495/08/03/1575/14 $2.00

doi: 10.1529/biophysj.107.119651

However, given that a few key residues known as the folding nucleus provide a significant driving force in the folding of a protein (26–28), we hypothesized that the way in which this nucleus is distributed between fragments determines reassembly efficiency of split proteins. In support of this notion, it has been observed that introduction of residues into the folding nucleus that lower its stability can dramatically slow the folding process (29).

To test our hypothesis, we have developed an on-lattice minimalist coarse-grained protein model to address how the reassembly kinetics, thermodynamic stability, and folding mechanism of a lattice model protein are affected upon splitting. Specifically, we designed several two-fragment systems derived from a well characterized 48-mer that is known to follow a nucleation-driven folding mechanism (30). Each of these split 48-mers was analyzed to determine the extent to which the reassembly process was impacted by differential partitioning of the folding nucleus between the two fragments. Our results suggest that a balanced distribution of folding nuclei amino acids between protein fragments is essential for efficient reassembly; this result was corroborated by the behavior observed for the reassembly process of a second set of split proteins derived from a 64-mer model protein. Collectively, these results provide new insights into the thermodynamic and kinetic aspects underlying protein fragment complementation and should prove extremely useful in the forward design and engineering of new split proteins.

METHODS

Split protein models

To explore protein fragment complementation experimentally, three two-protein fragment systems (N-split, Mid-split, and C-split) were created by splitting a model 48-mer protein, namely 48-1 (TSKRQQPYPMSLGSPFIR-IPMIGPRPRMRLLILLMGYPKRGRSGGGLF) (31), in three different locations (Fig. 1). Folded structures and a detailed thermodynamic and kinetic characterization for the parental 48-1 model protein sequence can be found elsewhere (31–33). In the N-split case, the sequence was split near the N-terminus between amino acids 16 and 17, creating one 16-residue fragment and a second 32-residue fragment. In the Mid-split case, the sequence was split in the middle between residues 24 and 25, creating two equal-sized fragments. In the C-split case, the sequence was split C-terminally between residues 32 and 33, creating one 16-residue fragment and a second 32-residue fragment. The symmetry shared by the N- and C-split systems was created so that the two fragments in each system were of equal length (i.e., each system has one 16-mer and one 32-mer fragment). This was done to eliminate any

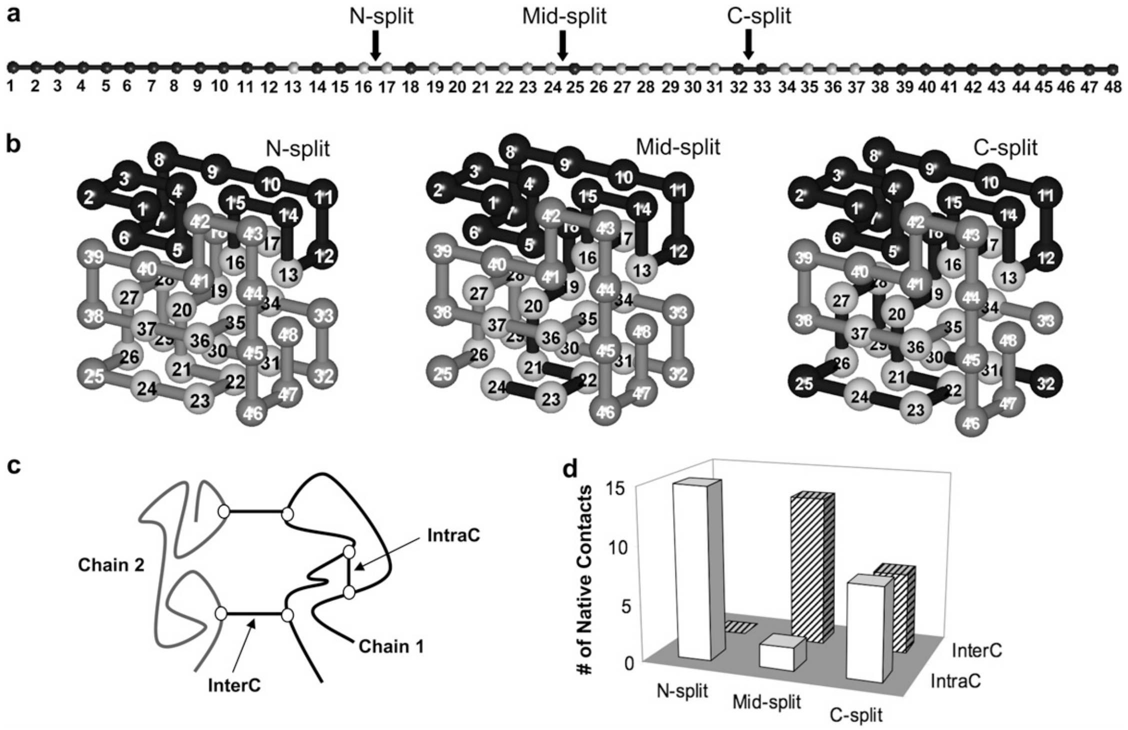

FIGURE 1 Construction of split protein systems. (*a*) Linear representation of the parent 48-mer sequence. Fragmentation sites for N-, Mid-, and C-split systems are indicated. Residues comprising the folding nucleus at the center of the folded structure are shown in light gray. (*b*) Schematic representation of folded structures for split systems. Amino acids or connecting bonds shown in black correspond to N-terminal fragment (chain 1); amino acids or connecting bonds shown in dark gray correspond to those in the C-terminal fragment (chain 2); amino acids shown in light gray correspond to the folding nucleus. (*c*) Schematic representation of interchain (InterC) contacts that involve interacting residues from both chains, and intrachain (IntraC) contacts that involve interacting residues within the same chain. (*d*) InterC and IntraC contacts formed by critical core residues upon protein fragmentation. Most probable native contacts found in the transition state (TS) ensemble (i.e., the folding core/nuclei) for folding of the 48-mer sequence at $T_f = 0.27$. Data obtained from Borrero and Escobedo (34).

effect on folding due to variations in chain size since it was unclear at the outset how this might impact reassembly.

Modeling folding for two protein fragments

To model the folding process, we adopted an on-lattice minimalist protein model in which the configuration of each protein chain evolves according to a canonical Monte Carlo (MC) algorithm (34). Briefly, space was discretized into a three-dimensional cubic lattice. Proteins were represented as self-avoiding chains, where each bead represents an amino acid with the bonds between the amino acids having uniform length equal to the lattice spacing (σ). Amino acid interactions were simulated by a Miyazawa-Jerningan contact energy potential (35) that takes into account implicit solvent effects and side-chain character. Conformational sampling was performed through a set of MC moves based on the Verdier-Stockmayer algorithm that mimics the diffusive movement of the amino acids during the folding process and includes 1), tail moves of one of the end beads to one of the available four neighboring sites; 2), corner flips for beads characterized by a right angle between directions to both contour neighbors; and 3), crankshaft moves of bead pairs located at the bottom of a U-turn (36). Relative to Verdier-Stockmayer moves, translation of a randomly selected chain was attempted after each MC step with a priori probability $\leq 10^{-4}$, consisting of adding either $+1$ or -1 (randomly chosen) to a random axis coordinate of all segment positions. Although this choice of translational move probability has no impact on thermodynamic averages, it affects the apparent kinetic dynamics of the system; for this reason, we only considered relative comparisons of real time kinetics between simulated dynamics for the 48-mer and the split systems (37).

Characterization of the folded state

To capture the specific chain topology of the folded state, two main parameters were used: the native energy (E_{nat}), which records the sum of the energies of all interresidue contacts, and the similarity parameter (Q), which represents the number of native contacts formed divided by the total number of native contacts that describes the folded structure of each system (38). According to this convention, $Q = 1$ represents the native (folded) conformation and $Q = 0$ represents the highly extended (unfolded) protein. As previously reported, the configuration corresponding to the folded state of the 48-mer structure was distinguished among all other visited configurations by the formation of 57 native contacts and a minimum energy value of $-20.24 k_{\mathrm{B}}T$ (31–33).

Spatial restriction

To make the association event more likely to occur without unduly constraining the conformations of the individual chains, space was restricted to a cubic box of 12σ length units, corresponding to a volume fraction of chains of ~3%. However, given the small size ($3 \times 4 \times 4$) of the folded structure, the small size of each chain (16–32 residues), and the small number of chains involved (two), this spatial restriction was closer to a diluted regime, since the chains had plenty of free space to move. It is also worth noting that by encaging the system, we essentially disregarded the diffusion process that needs to occur before the two chains come near each other; instead, we consider a restricted open space where the local environment was crowded enough to allow for interchain interactions but precluded the chains separating to an infinite distance.

Additional simulation parameters

To collect kinetic data, simulations were run up to the point where the native structure of the system was observed for the first time, and this time was recorded as the folding time. In the case where no folding was observed, simulations were run for a maximum of 5×10^8 MC steps. Data from each simulation was obtained by taking the mean folding time (MFT) values over 500 independent runs in the canonical ensemble, each one starting from a different unfolded structure ($Q \leq 0.2$). Results were determined to be statistically invariant, since the data was not significantly affected when additional runs beyond 500 were included in each simulation.

Thermodynamic analysis

The thermodynamics of the single and multichain systems were studied by employing replica exchange MC (REMC) sampling (36,39) combined with the multihistogram reweighting method (MHR) (40). REMC was used to alleviate problems related to the sampling of a rugged free-energy landscape, in which the polypeptide chains could be temporarily trapped at low temperature. Protein folding was simulated by running several parallel replicas (M), each at a different temperature (T_{i}). The reduced temperature, T, was normalized by the reference temperature, T_{o}, such that $k_{\mathrm{B}}T_{\mathrm{o}}$ represented the energy unit pertinent to the system. Relative to Verdier-Stockmayer and translation moves, swap moves between systems of different temperatures were attempted after each MC step with a probability ≤ 0.05. In most calculations, the number of replicas was 9, with T ranging between 0.1 and 0.5. Details of the thermodynamic analysis are given in (32). By using the REMC-MHR method, data from all replicas were combined and analyzed, minimizing the error in the estimation of the density of state function [$\Omega(E)$] and facilitating the calculation of thermodynamic quantities over a wide range of temperatures, such as the specific heat (C_{v}) via Eq. 1, and free energy via Eq. 2:

$$C_{\mathrm{V}}(T) = \frac{\langle E^2 \rangle_{\mathrm{T}} - \langle E \rangle_{\mathrm{T}}^2}{k_{\mathrm{B}}T^2}; \quad (1)$$

$$A_{\mathrm{E}}(E,T) = E - TS = -k_{\mathrm{B}}T(\ln(P(E,T) - \ln(Z(T)). \quad (2)$$

Here, E represents the energy, k_{B} is Boltzmann's constant, T is the temperature, S is the entropy, the partition function $Z(T) = \Sigma_{\mathrm{E}}\, \Omega(E)\exp(-E/k_{\mathrm{B}}T)$, and the Boltzmann distribution of states $P(E,T) = \Omega(E)\exp(-E/k_{\mathrm{B}}T)/Z(T)$.

RESULTS

Design of protein fragments

For this study, we chose the model 48-mer protein, 48-1, because its thermodynamic behavior, folding pathway, and transition state have been characterized in detail (31–33). The 48-1 sequence was originally designed by Shakhnovich and co-workers to model a well designed sequence that exhibits a stable, fast-folding structure and an all-or-none transition between clearly distinguishable native and unfolded states (31). To generate split lattice model proteins, we dissected the 48-1 sequence at three positions: between residues 16 and 17 (N-split), 24 and 25 (Mid-split), and 32 and 33 (C-split) (Fig. 1 *a*). The minimum-energy folded structure recovered from a large MC simulation for each of the N-, Mid-, and C-split systems (Fig. 1 *b*) was identical to that reached by the unsplit 48-1 chain (data not shown). However, whereas unsplit 48-1 was characterized by 57 native contacts, the folded state for all split cases was characterized by 58 native contacts, since the additional contact lost upon the excision of the full chain needed to reform between the last amino acid of the first fragment and the first amino acid of the second fragment. Additionally, as a result of this new native contact, the energy values for the N-, Mid-, and C-split systems were -20.43, -20.65, and $-20.62k_BT$, respectively, compared to $-20.24k_BT$ for the unsplit 48-mer. It is also worth noting that the split sites for the N-, Mid-, and C-split systems were involved in five, three, and two total native contacts (including the split pair), respectively, that contributed locally to ~8%, 5%, and 4%, respectively, of the total native energy.

More recently, it was shown that the 48-1 protein folds according to a classical nucleation mechanism, whereby a core of native contacts forms at an early stage of the process and causes the protein to rapidly collapse to more compact nativelike conformations that lead to the fast rearrangement of its residues into the final folded structure (34). These same authors reported that the nucleus was composed of several mostly hydrophobic amino acids that have >60% probability of forming native contacts in the transition-state intermediates; these residues (residues 13, 16, 17, 19–24, 26–31, and 34–47 in Fig. 1 *a*) form a core at the center of the folded structure. It is important to note that in the Mid- and C-split cases, folding nuclei residues are well distributed between fragments and participate in a significant number of interchain native contacts (InterC) as seen in Fig. 1, *c* and *d*. In contrast, for the N-split case, the folding nuclei residues are disproportionately distributed between fragments and none of these are involved in interchain native contacts (Fig. 1, *c* and *d*).

Thermal stability is affected by protein fragmentation and by choice of the split site

The effect of splitting on thermodynamics was studied by determining the transition temperature (T_{max}) for the unsplit 48-1 and each multichain system. A plot of heat capacity as a function of temperature revealed a single, strong peak corresponding to the folding temperature (T_{max}) for the 48-1, N-, Mid-, and C-split systems (Fig. 2), indicating a single-phase conformational transition. Relative to the single 48-mer chain, all of the two-fragment systems exhibited lower folding temperatures. Normalized transition temperatures were found to be $T_{max}/T_f = 1$ for the 48-mer, $T_{max}/T_f = 0.956$ for the C-split system, $T_{max}/T_f = 0.937$ for the Mid-split system, and $T_{max}/T_f = 0.926$ for the N-split system. Thus, whereas the unsplit 48-mer remained stable at a higher temperature, thermal denaturation occurred at lower temperatures when protein folding was reconstituted from multiple fragments. These data also suggest that thermal denaturation was dependent on the choice of split site, as evidenced by the difference in folding temperatures between the entirely symmetric N- and C-split systems.

Whereas we did not explicitly test the effect of protein concentration in this study, the decrease in thermal stability observed in the context of split fragments was consistent with the earlier observation that folding temperature decreased as the concentration of protein chains increased in a system designed to mimic protein aggregation (41). The observed decrease here was related to both 1), an increase in the frequency with which the protein's configurational energies were close to that of the unfolded state ($Q \approx 0$); and 2), a decrease in the frequency with which the multichain system explored nativelike configurations ($Q \approx 1.0$) during the folding process. For a more detailed analysis, let us assume a pseudoreaction of the following form for the unsplit 48-mer:

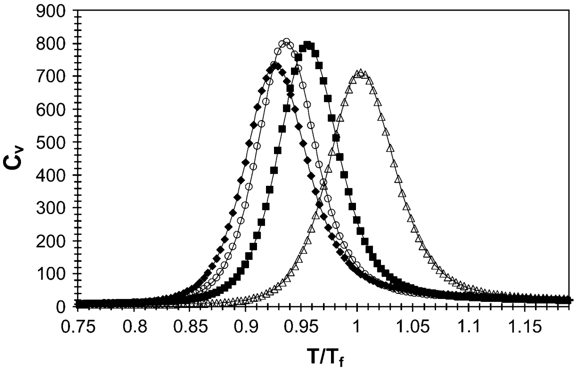

FIGURE 2 Thermodynamic analysis of single- and multichain systems. Heat capacities for the 48-1 (Δ), N-split (♦), Mid-split (O), and C-split (■) proteins as a function of temperature within the scale of the energy potential implemented in our model. Transition temperature (T_{max}), also referred to as the protein's folding temperature, was defined as the temperature at which the heat capacity exhibits a maximum for each system. All temperatures were normalized by 0.27, the folding temperature of the unsplit 48-mer (T_f). Thermodynamic simulations were performed for 5E9 MC steps (a long time relative to folding times).

$$\text{Unfolded} \xrightarrow{\text{48mer}} \text{Folded}; \tag{3}$$

and, for the two-fragment split systems,

$$\text{Unfolded-1} + \text{Unfolded-2} \xrightarrow{\text{Split}} \text{Folded}(1+2). \tag{4}$$

The increase in the number of available nonnative configurations stems from the fact that two chains have more freedom to explore the conformational space separately, and this increases the entropy of the unfolded state. The total entropy change upon folding involved in both the unsplit 48-mer ($\Delta S_{48\text{-mer}}$) and the split systems (ΔS_{split}) has two main contributions, one due to the reduction of conformational entropy (S^{Conf}) and the other due to a reduction of translational entropy (S^{Trans}). Using Flory's lattice model to count chain conformations, it can be shown that S^{Conf}can be approximated as

$$S^{Conf}/k_B = N[(1 - 1/\rho)\ln(1 - \rho) - 1], \tag{5}$$

where k_B is Boltzmann's constant, N is the number of segments (i.e., amino acids) in a chain, and ρ is the segment density (i.e., the number of amino acids within the volume occupied by the chain) (36). Since the folded state can be taken as a maximally collapsed state ($\rho \rightarrow 1$) and the unfolded state as an open conformation ($\rho \rightarrow 0$), a first approximation for the $\Delta S^{\mathrm{Conf}}$is given by

$$\Delta S^{\mathrm{Conf}}_{\mathrm{48mer}} = -Nk_{\mathrm{B}} \tag{6}$$

and

$$\Delta S^{\mathrm{Conf}}_{\mathrm{split}} = -(N_1 + N_2)k_{\mathrm{B}}, \tag{7}$$

where N_1 and N_2 represent the number of amino acids in each of the two fragments in the split system. Since the length of the unsplit system is $N = N_1 + N_2$, it follows that the unsplit 48-mer and all the derived split systems entail a roughly similarΔS^{Conf}. The second entropic contribution S^{Trans} for an ideal molecule (lacking interactions with other molecules) is given by

$$S^{\mathrm{Trans}}/k_{\mathrm{B}} = n[3/2 + \ln(V/n)], \tag{8}$$

where n is the number of molecules and V is the volume accessible to them (e.g., in units of molecular volume) (36). According to Eq. 8, the unsplit 48-mer folding process entails no change of translational entropy ($\Delta S^{\mathrm{Trans}}_{\mathrm{48mer}} = 0$), since the number of molecules does not change upon folding ($\Delta n = 0$) and the entropy is independent of the chain's center of mass. In contrast, the change of translational entropy upon folding for the split processes is given by (with $\Delta n = -1$ and assuming $V \gg 1$):

$$\Delta S^{\mathrm{Trans}}_{\mathrm{split}} = -k_{\mathrm{B}}(\xi + \ln V), \tag{9}$$

where ξ is a positive constant whose precise value is not important. Hence, when calculating the total entropic difference upon folding between the split and unsplit processes (i.e., Eq. 9 + Eq. 7 − Eq. 6), a change of$\Delta S^{\mathrm{Trans}}_{\mathrm{split}}$ is obtained; the fact that this change is always negative indicates that, relative to the folding process of the unsplit 48-mer, the folding process of the split systems results in an overall unfavorable entropic change (i.e.,$-k_{\mathrm{B}}(\xi + \ln V)$).

In addition to the entropic differences between the unsplit and split systems, the enthalpy change associated with the folding process (computed from the difference between the average configurational energy of the folded (E^{F}) and unfolded (E^{U}) states) of the split systems is also unfavorable relative to the enthalpy change associated with the folding of the single 48-mer chain. In this case, $\Delta E = E^{\mathrm{F}} - E^{\mathrm{U}}$ increases for the split proteins because the energy of the unfolded state decreases with the number of protein fragments. The lower energy of the unfolded state in split systems can be rationalized by the fact that protein fragmentation allows more freedom for some favorable contacts to form that are not able to form in the unsplit 48-mer (where all amino acids are connected). As shown in Fig. 3, the multichain system can sample configurations around the unfolded state for a range of energies that are not available for the unsplit system. In these plots, free energy landscapes for the unsplit and split chains are projected over the plane of native energy and the fractional nativeness. Note that the configurational energy refers to the total energy of the system (i.e., sum of the configurational energy for chain 1 and chain 2, and that between the two chains). If we assume that the folded state has essentially the same average energy (E^{F}) for the unsplit 48-mer and split systems, the difference in energy between these two processes is always positive, as shown below:

$$\Delta E_{\mathrm{split}} - \Delta E_{\mathrm{48mer}} = E^{\mathrm{U}}_{\mathrm{48mer}} - E^{\mathrm{U}}_{\mathrm{split}} > 0. \tag{10}$$

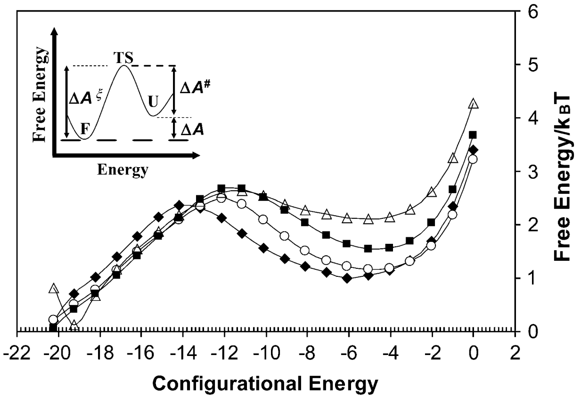

FIGURE 3 Free energy (ΔA) versus configurational energy at $T = 0.25$ for: 48-1 (Δ), N-split (♦), Mid-split (○), and C-split (■) systems. The inset gives a schematic diagram of the free energy of the native ($\Delta A^{\xi} = \mathrm{A}^{\mathrm{TS}} - A^{\mathrm{F}}$) and unfolded state ($\Delta A^{\#} = A^{\mathrm{TS}} - A^{\mathrm{U}}$), and the free energy of stabilization ($\Delta A = \Delta A^{\#} - \Delta A^{\xi}$). The folded state (F) is defined by the minimum found at the lowest configurational energy, the transition state (TS) is defined by the maximum (peak) of the free-energy curve, and the unfolded state (U) is defined by the minimum found at the highest configurational energy.

Collectively, this analysis indicates that the reduced thermodynamic stability of the native state in the split systems arises from two factors: unfavorable enthalpic and unfavorable entropic contributions.

The thermodynamic destabilization of the assembled split chains is also reflected by their higher free energies (ΔA) relative to the free energies observed in the case of the unsplit 48-mer (Fig. 3). ΔA is defined as the difference in free energy change between the folded state (A^{F}) and the unfolded state (A^{U}), i.e., $\Delta A = A^{\mathrm{F}} - A^{\mathrm{U}}$. Using Eq. 10, the difference in free energy changes between the unsplit 48-mer and the split-chain systems can be found by Eq. 11 (i.e., Eq. 13 − Eq. 12):

$$\Delta A_{\mathrm{split}} - \Delta A_{\mathrm{48mer}} = E^{\mathrm{U}}_{\mathrm{48mer}} - E^{\mathrm{U}}_{\mathrm{split}} + k_{\mathrm{B}}T(\xi + \ln V), \tag{11}$$

where

$$\Delta A_{\mathrm{48mer}} = E^{\mathrm{F}} - E^{\mathrm{U}}_{\mathrm{48mer}} + k_{\mathrm{B}}T(N_1 + N_2), \tag{12}$$

and

$$\Delta A_{\mathrm{split}} = E^{\mathrm{F}} - E^{\mathrm{U}}_{\mathrm{split}} + k_{\mathrm{B}}T(N_1 + N_2) + k_{\mathrm{B}}T(\xi + \ln V). \tag{13}$$

Since we have already argued that all terms on the righthand side of the equation are positive (see Eqs. 9 and 10), the difference in free energy between the 48-mer and the split protein systems (as calculated by Eq. 11) is always positive as the system goes from the unfolded to the folded state. Importantly, the fact that $(\Delta A_{\mathrm{split}}) > (\Delta A_{\mathrm{48mer}})$ indicates that the folding of any split-chain system will have a smaller thermodynamic driving force than its corresponding unsplit system. The relevance of the cage volume (V) on multichain folding can also be appreciated from this simple thermodynamic model.

Kinetics of protein reassembly is sensitive to the split site

To determine the effect of temperature on the relative folding kinetics of the different split protein systems, we calculated the mean folding time for the 48-mer and N-, Mid-, and C-split systems over a wide range of temperatures. The optimum temperature (T_{opt}), defined as the temperature at which a given system folds fastest, was ~0.23 for the 48-mer, 0.22 for N-split, 0.23 for Mid-split, and 0.22 for C-split (Fig. 4). The MFTs for the N- and Mid-split proteins were approximately three and two times slower, respectively, than that of the 48-mer at their corresponding T_{opt}s (Fig. 4). Importantly, the total number of independent runs where the native structure formed within the maximum simulation time (5×10^8 MC steps) was 500 out of 500, or 100%, for each system. This percentage was defined as the folding frequency (FF). The apparent folding rate (AFR), defined as the ratio of FF to MFT at T_{opt}, was determined to be 1.92×10^{-5} for N-split, 3.57×10^{-5} for Mid-split, and 6.37×10^{-5} for C-split.

The slower kinetics of fragment reassembly, relative to the folding of a single chain, is not entirely surprising. Intuitively, this could be partially reasoned by the fact that all the residues that need to come into contact to form the folded structure in a single chain are in closer proximity by virtue of their interconnectivity; this is strikingly different from the case of two unconnected chains, where residues that have to associate to enable the formation of native contacts can move independently in space. Thermodynamically, the increase in folding times for the split fragments relative to the folding time of the single 48-mer chain is also not surprising, since it can be argued that the reassembly of split fragments (represented by Eq. 4) has a larger free-energy barrier ($\Delta A^{\#} = A^{\mathrm{TS}} - A^{\mathrm{U}}$) and thus should be slower than the folding process for the unsplit 48-mer (represented by Eq. 3). This conjecture can be reached by assuming that the folding ''transition state'' (TS) is roughly independent of whether or not the protein is split, the ''folded'' state (F) on the righthand sides of Eqs. 3 and 4 can be replaced by the TS. Although the assumption of TS isomorphism is not generally justified, since the TS should depend on the location of the splitting site, it is sensible to expect that the relative decrease of the free energy of the unfolded state (embodied by Eq. 13) in any two-chain system will also tend to increase the barrier to folding (for the same underlying physical reasons).

Two aspects of the kinetic data shown in Fig. 4 are unexpected and intriguing: 1), the observation that a much smaller change in folding kinetics exists between the 48-mer and the C-split system (relative to the 48-mer and the other split systems), to the extent that there is no significant change in the folding times of these two systems at temperatures neighboring their respective T_{opt}s; and 2), the observation that at T_{opt} the N-split folds 46% slower than the Mid-split and 70% slower than the C-split, despite the complete symmetry of these two systems. These trends prevailed over most of the temperature range tested for each system. It is also worth noting that the fragmentation itself did not dramatically retard folding in the case of the C-split system. This can best be attributed to the spatial constrictions that were placed on this moderately confined system (3-D cage of size 12σ), where a crowded environment relative to open space was created to ensure association between the different fragments. Note that it has been previously shown that, relative to folding in open space, the folding kinetics of this particular unsplit 48-mer remain unchanged when confined within a cubic box of size $>10\sigma$ unit length (32,33).

The differences in folding kinetics can be rationalized thermodynamically by comparing the differences in free-energy barriers observed between the 48-mer and the different split proteins. For instance, the similarity in folding kinetics between the unsplit 48-mer and the C-split system is reflected in Fig. 3. These data show that, although $\Delta A^{\#}$ is larger for the C-split than for the unsplit system, these two systems display approximately the same TS dividing surface. Likewise, the much slower folding kinetics between the Mid-split and especially the N-split system relative to the unsplit 48-mer is reflected by the displacement of the TS toward the

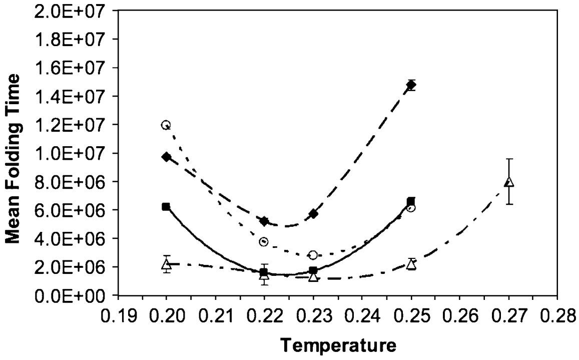

FIGURE 4 Kinetic analysis of model proteins. Mean folding time (MFT) plotted over a range of temperatures for the 48-mer 48-1 (△), N-split (♦), Mid-split (○), and C-split (■) proteins. The MFT value corresponds to the MC step at which the folded structure was first observed. Each data point was obtained from an average of 500 simulations. Error values were estimated by finding the difference between the mean folding time calculated from the first 250 simulations and the mean folding time obtained for the last 250 simulations; this value was then divided by 2. The errors are within the symbol size.

folded state (i.e., toward states of lower configurational energies, where it is more difficult to be accessed). The shift in the transition-state dividing surface observed for the N- and Mid-split systems, but not the C-split, indicates that the reassembly of these systems takes place via a different folding mechanism that appears to be slower. Collectively, our kinetic data and thermodynamic analysis of free energies suggest that in this confined system, the degree of retardation observed as a result of having two separate fragments is modulated by the location of the splitting site with respect to the folding nucleus.

The roughness on the free energy folding landscapes depends on the split site

To further explore the differences underlying the observed trends in MFTs, we plotted the free-energy landscape of the 48-mer, N-, Mid-, and C-split systems at their respective $T_{\max}$ as a function of the total contact energy and the similarity parameter Q (Fig. 5). In the case of the split-fragment systems, the parameter Q included native contacts that formed within the same chain (intrachain) as well as those formed between different chains (interchain) (Fig. 1, *IntraC* and *InterC*, respectively). The free energy was obtained from Eq. 2. Consistent with previous work, the 48-mer exhibited two free-energy minima corresponding to the unfolded (high energy, $Q \approx 0$) and folded (low energy, $Q \approx 1$) states that were connected by a relative narrow passage wherein the transition state was identified as a saddle point (Fig. 5 *a*). The narrowness of the connecting region between the unfolded and folded states was characteristic of well designed proteins that exhibit a minimum number of misfolded (i.e., low-energy, low-Q-structure) states (42).

The fact that the same lowest energy configuration state was observed in all the landscapes confirmed that all systems shared the same folded state (Table 1). Moreover, since the additional contact observed in the split systems was favorable, the total configurational energy of these systems decreased with respect to the unsplit 48-mer. It is also important to stress that this folded state remained unique and was only achieved by the reassembly of the two chains; this is implicitly suggested by Fig. 5, *b–d*, where only one low-energy state with a large number of native contacts was observed. The absence of multiple local energy minima in a region of a large number of native contacts supports the observation that single fragments by themselves remained unstructured and high in energy relative to the state they formed upon assembly. These differences separated these landscapes from those observed in a multichain aggregation system (41), where the appearance of low-energy/high-Q states suggested that each chain folded independently and that the formation of interprotein contacts only inhibited their separate folding process and resulted in aggregated, high-energy/low-Q states.

One striking difference observed in the folding landscape of the 48-mer (Fig. 5 *a*) when compared to the split proteins (Fig. 5, *b–d*) was the spread of the free-energy minima region neighboring the unfolded state across a wider range of low Q values, closer to the transition state region of the parent 48-mer protein. This observation was significant, since the

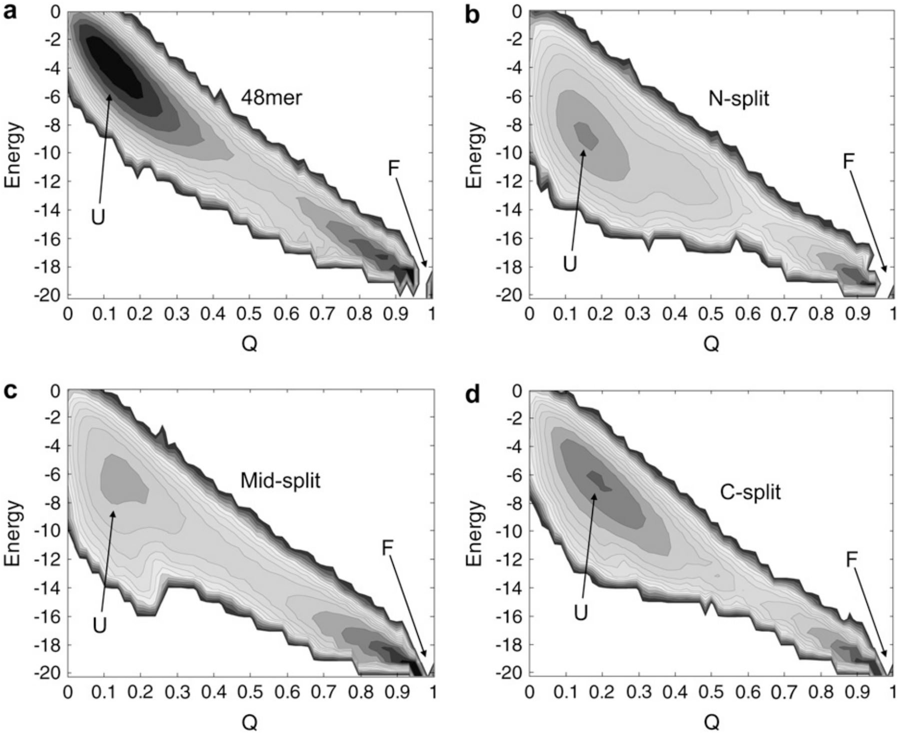

FIGURE 5 Free-energy landscape for all protein systems at $T_{\max}$. Contour plot of the free-energy landscape of the 48-1-mer (*a*), N-split (*b*), Mid-split (*c*), and C-split (*d*) at the $T_{\max}$ for each protein. The lowest elevations are indicated by arrows and appear as darkly shaded regions in the upper left (unfolded basin, *U*) and lower right (native-state basin, *F*) of each panel. Q values (*x* axis) represent the fraction of native contacts, calculated as the number of native contacts formed divided by the total number of native contacts for each sequence (i.e., 57 total contacts for the 48-mer and 58 native contacts for the split proteins). Energy values (*y* axis) represent the total configurational energy for the system (i.e., the sum of all energies for contacts within chain 1, chain 2, and between the two chains). The folded and unfolded states are represented by the two minima at $Q = 1.0$ and $Q \approx 0.1$, respectively. Simulations were performed for a total number of 5×10^9 MC steps at equilibrium.

TABLE 1 Native contact pairs observed in the 48-mer folded structure

NC pair code	*NC* pair (*i*, *j*)		*NC* pair code	*NC* pair (*i*, *j*)	
	i	*j*		*i*	*j*
1	T1	R4	31*	I19	M28
2	T1	R40	32*	I19	L30
3	T1	R42	33*	I19	L34
4	S2	K39	34*	P20	R27
5	K3	Q6	35*	P20	M35
6	K3	Y8	36*	P20	Y37
7	R4	P9	37*	M21	P24
8	R4	P15	38*	M21	P26
9	Q5	F16	39*	M21	L30
10	Q5	R18	40	I22	L31
11	Q5	P20	41*	I22	M35
12	Q5	R40	42	I22	L47
13	Q6	R27	43*	G23	G36
14	Q6	K39	44	G23	G46
15	P7	R18	45*	P24	Y37
16	P7	M28	46	R25	P38
17	P9	R18	47*	P26	R29
18	M10	P15	48	R27	P38
19	M10	I17	49	L31	L34
20	S11	S14	50	I32	L47
21	L12	I17	51	L33	F48
22	L12	L33	52	M35	F48
23*	G13	F16	53	G36	G41
24	G13	G44	54	G36	G45
25	G13	F48	55	Y37	R40
26	S14	S43	56	G41	G44
27	P15	R42	57	G45	F48
28	F16	M35	58[†] N-split	F16	I17
29	F16	G41	Mid-split	P24	R25
30*	I17	L34	C-split	I32	L33

All native contacts (NC) found in the folded structure are listed. The pair code numbers for NCs (*left column*) correspond to the same numbering (1–58) used in Fig. 7 to represent NCs. The NC pair (*i*, *j*) describes an interaction between amino acids *i* and *j*, where *i* and *j* indicate the type of amino acid and its position in the 48-mer sequence (e.g., pair 1 describes an interaction between the threonine found at position 1 (T1) and the arginine found at position 4 (R4) in the unsplit 48-mer sequence).

*NCs that form the critical folding nuclei (listed in Table 2).

[†] One extra contact describes the folded structure of the split systems as a result of the additional link that needs to form at the site of fragmentation.

extent to which this low-energy, misfolded (low energy/low *Q*) region was amplified directly correlated with the retardation observed in the kinetics of the reassembly process. That is, whereas the free-energy landscape of the 48-mer did not change significantly when splitting the protein C-terminally (Fig. 5, *a* versus *d*), a much more diffusive (i.e., broad and rough) passage from the unfolded to the folded state resulted when splitting the 48-mer near its N-terminus (compare Fig. 5, *a* and *b*).These data suggest that the efficiency of the reassembly process was decreased by the entrapment of protein fragments in misfolded configurations. Given that slower folding kinetics and a diffusive free-energy landscape were observed for the N-split relative to the C-split system, we hypothesized that the shared distribution of critical core residues between the two fragments is essential for efficient reassembly. This hypothesis is supported by the observation that the distribution pattern of critical core residues is the primary difference between the N-split and C-split fragments.

Productivity of interchain interaction depends on split site

The inefficiency in folding observed for the N-split relative to other systems could have resulted from lack of association between the two fragments (i.e., the fragments never came together) or, if they did associate, from an inability of the fragments to form productive interactions. Since the parameter *Q* includes both interchain and intrachain native contacts, the free energy landscapes shown in Fig. 5 do not distinguish between misfolded configurations caused by unproductive interactions between the two fragments and those caused from unproductive interactions among individual fragments. To decouple this effect, we plotted contours of the number of interchain contacts as a function of the similarity parameter, *Q*, for the N-, Mid-, and C-split systems at their respective $T_{\max}$ (see Fig. 1 in Supplementary Material). Two highly populated regions were observed in these landscapes. The first region, representing a large number of interchain contacts neighboring the folded state (high InterC, high *Q*), confirmed that access to the folded state was highly dependent on associations between chains. The second region represented a significant (but not high) number of interchain contacts neighboring the unfolded state (mid-InterC, low *Q*) and was much more populated for the N-split than for the Mid- and C-split systems. This observation suggests that although associations between fragments occurred for all the systems, the occurrence of these in the N-split case was less likely to result in productive interactions that would lead to the folded state. Taken together, these data support the notion that the efficiency of protein reassembly depends to a great extent on the site at which the protein is split.

A shared critical nucleus "glues" fragments productively during reassembly

Given that the formation of the critical nucleus is key for folding efficiency in the case of a classical nucleation folding mechanism, as is the case for the 48-mer (32), we next analyzed how the dissection of amino acids in the nucleus upon protein fragmentation affected reassembly and folding. Specifically, we plotted landscapes of the critical core residues (Table 2 and Fig. 1 *d*) as a function of the total number of native contacts (*Q*). In the N-split case, a region with a high number of critical contacts and a low number of total native contacts was observed (Fig. 6 *a*), but not in the case of the Mid- or C-split proteins (Fig. 6, *b* and *c*). These data indicate that the more difficult transition to the folded

TABLE 2 Distribution of native contacts forming the folding nuclei in split proteins

NC pairs (*i*, *j*)		N-split		Mid-split		C-split	
i	*j*	InterC	IntraC	InterC	IntraC	InterC	IntraC
P20	M35		2	X		X	
M21	P24		2		1		1
I19	L34		2	X		X	
P20	Y37		2	X		X	
I19	L30		2	X			1
I22	M35		2	X		X	
G23	G36		2	X		X	
P20	R27		2	X			1
M21	L30		2	X			1
M21	P26		2	X			1
I19	M28		2	X			1
G13	F16		1		1		1
P24	Y37		2	X		X	
I17	L34		2	X		X	
P26	R29		2		2		1

Most probable native contacts found in the transition-state ensemble (i.e., the folding core/nuclei) for folding of the 48-mer sequence at $T_{\mathrm{f}} = 0.27$ are listed in order of decreasing probability. The NC pair (*i*, *j*) describes the interacting pair, where *i* and *j* entries indicate the type of amino acid and its position in the unsplit 48-mer sequence. The distribution of these contacts in all the split proteins is marked as follows: interchain contacts (InterC), which involve interacting residues from both chains are marked by an ''X'' and intrachain contacts (IntraC), which involve interacting residues within the same chain, are marked by a number (1 or 2) that specifies the fragment where the interaction takes place. Fragments 1 and 2 for each system correspond to those illustrated in Fig. 1.

state observed for the N-split protein stems from the formation of the full core in a single chain that trapped the system in a region of highly misfolded states. Further analysis of folding ''snapshots'' of the N-split system during a typical folding trajectory suggests that intrachain formation of the core leads to preassembly of the largest fragment (chain 2) into a semistable structure that prevents the efficient incorporation of the smallest chain (chain 1) (Fig. 6 *a*). This type of isolated preassembled structure was clearly observed in the snapshots (Fig. 6 *a*, *i* and *ii*), where these chains exhibited minimum association with each other. In stark contrast, the shared formation of the core between the Mid- and C-split systems resulted in transition-state structures of highly interacting fragments that more readily formed the rest of the native contacts, leading to efficient assembly of the folded structure (Fig. 6, *b* and *c*). However, although these structural configurations were characterized by the formation of interchain native contacts, the part of the fragments that was away from the contact point between the two chains remained highly extended. The structural patterns reflected in these snapshots were repeatedly observed throughout the 10–15 sets of data that we analyzed for each system (data not shown).

The simple thermodynamic model presented above (see Eqs. 5–9) was used to rationalize the differences in behavior between the case where one of the two chains preassembles (such as the N-split case) and the case where both chains exhibit more cooperative folding behavior (such as the C-split case). For this analysis, we assume that the ''unfolded'' state is the one in which the two chain fragments have already collapsed or associated, if strongly inclined to do so. Based on the typical snapshots analyzed for the folding trajectory of the N-split case (Fig. 6 *a*), we assume that in the unfolded state, chain 1 (the small chain) has an open conformation (with $\Delta S^{\mathrm{Conf}}_{\mathrm{split}} = -N_1 k_{\mathrm{B}}$), chain 2 is prefolded (with $\Delta S^{\mathrm{Conf}} \rightarrow 0$), and the two chains tend to be separate (with $\Delta S^{\mathrm{Trans}}_{\mathrm{split}} = -k_{\mathrm{B}}(\xi + \ln V)$); in this case, the total (conformational and translational) entropy can be described as $\Delta S_{\mathrm{N\text{-}split}}/k_{\mathrm{B}} = -N_1 - \xi - \ln V$. Consistent with Fig. 6, for the C-split case, we assume that in the unfolded state both chains are not collapsed but tend to be associated (with $\Delta S^{\mathrm{Trans}} \rightarrow 0$); in this case, the total entropy is purely conformational and can be described as: $\Delta S_{\mathrm{C\text{-}split}} = -(N_1 + N_2)k_{\mathrm{B}}$.

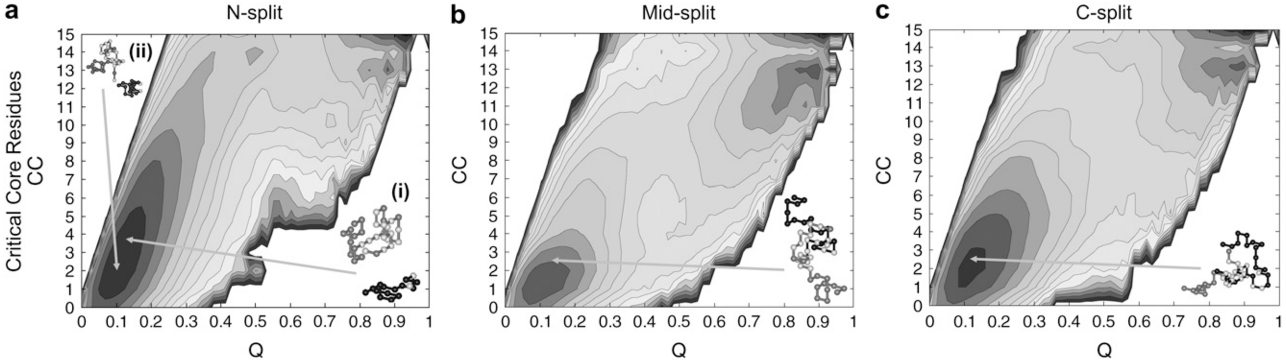

FIGURE 6 Free-energy landscapes for native contacts forming the critical core. Contour plot of the free-energy landscapes for the critical core residues of the N-split (*a*), Mid-split (*b*), and C-split (*c*) proteins at T_{max}. The lowest elevations appear as darkly shaded regions in the lower left (unfolded basin (*arrows*)) and upper right (native-state basin) of each panel. The folded and unfolded states are represented by the two minima at $Q = 1.0$ and $Q \approx 0.1$, respectively. Insets are snapshots depicting typical configurations observed for each case before the formation of the folded structure (i.e., $Q < 0.5$). Shading of the chains is as described for Fig. 1.

Given these expressions for entropy, the free-energy changes upon folding, for the N-split and C-split cases, can be described as

$$\Delta A_{\text{N-split}} = E^{\text{F}} - E^{\text{U}}_{\text{N-split}} + k_{\text{B}}TN_1 + k_{\text{B}}T(\xi + \ln V) \qquad (14)$$

and

$$\Delta A_{\text{C-split}} = E^{\text{F}} - E^{\text{U}}_{\text{C-split}} + k_{\text{B}}T(N_1 + N_2), \qquad (15)$$

respectively, so that the difference between these two free-energy changes is

$$\Delta A_{\text{N-split}} - \Delta A_{\text{C-split}} = (E^{\text{U}}_{\text{C-split}} - E^{\text{U}}_{\text{N-split}}) + kT(\xi + \ln V) - kTN_2. \qquad (16)$$

Assuming that the unfolded N-split protein has stronger (more negative) energetic interactions than the unfolded C-split protein, then $E^{\text{U}}_{\text{C-split}} > E^{\text{U}}_{\text{N-split}}$; note that this result is consistent with Fig. 3, where we observed that average unfolded-state configurational energies were lower in the case of the N-split than in the case of the C-split system. Additionally, since our simulation results showed that the folded N-split protein was less stable than the folded C-split protein, we conclude that $\Delta A_{\text{N-split}} > \Delta A_{\text{C-split}}$. Based on these results, the righthand side of Eq. 16 must be positive. In this case, it appears that the first two (positive) terms in the lefthand side of Eq. 16 dominate, so that$\Delta A_{\text{N-split}} > \Delta A_{\text{C-split}}$. It is important to note that this result indicates that the driving force for folding is smaller for the N-split system than for the C-split system. Note, however, the nontrivial interplay of the interactions: 1), the prefolding of a chain fragment favors folding on entropic grounds (since the unfolded states start at lower entropies, e.g., more ordered) but disfavors folding on energetic grounds (since unfolded states are found at lower energies, e.g., closer to the folded state); and 2), the interchain association favors folding on entropic grounds (by reducing translational entropy) but may disfavor it if the associated (unfolded) states are found at very low energies.

A different folding mechanism emerges when core residues are not shared

To obtain insight into the mechanism by which the two fragments assemble, we examined the order in which all native contacts formed over 500 different folding trajectories for each split system. It was observed that the first native contacts to form (i.e., the ones with longer contact waiting time, τ_{f}) are those corresponding to the critical core (Fig. 7). Although a precise folding mechanism for the split fragments cannot be inferred by these results alone (i.e., specific transition states are not identified), these data indicate that 1), the same critical core (Table 1) of native contacts seen for the parent protein forms even in the cases when the protein is split; and 2), early formation of this set of native contacts is critical to the folding pathway of the split fragments.

Inspection of these data also suggests that the assembly mechanism of the N-split differs significantly from that of the Mid- and C-split cases. For instance, two separate stages were observed in the reassembly process of the N-split protein (Fig. 7 *a*). During the first stage (at longer τ_{f}), a set of critical native contacts preassembled in the longer chain (chain 2), whereas the smaller chain (chain 1) remained completely unincorporated (no interchain contacts were formed) and unfolded (no native contacts were observed). Then, during a later stage (at shorterτ_{f}), the folding process was completed when the smaller chain was incorporated into this already preassembled structure to form the rest of the native contacts. It is important to note that the coassembly stage did not take place until a long time (relative to the total folding time) after the folding process had started. A much

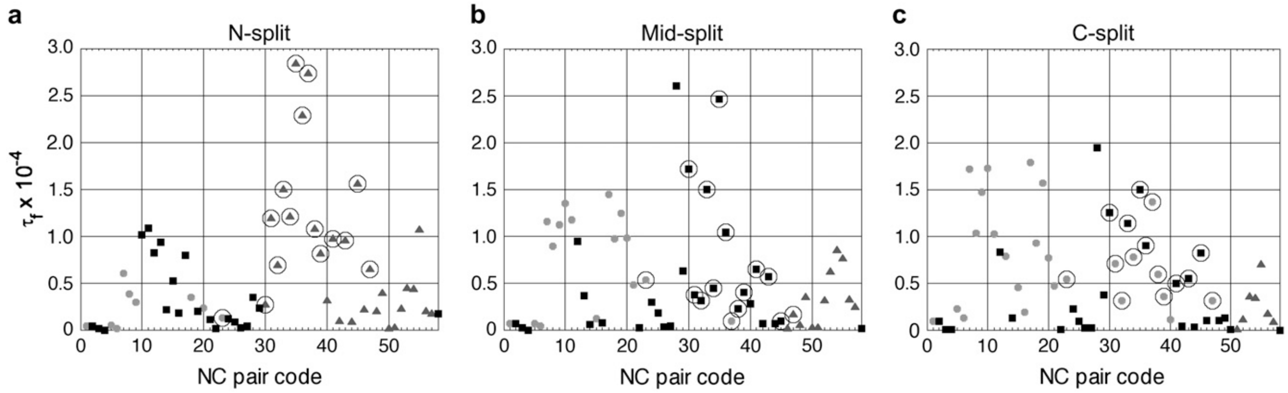

FIGURE 7 Kinetic evolution of native contact formation in split proteins. The contact waiting time (τ_{f}) represents the time a native contact has to wait until complete folding takes place for the N-split (*a*), Mid-split (*b*), and C-split (*c*) proteins. τ_{f} was normalized by the total folding time (MC step) for each protein and τ_{f} was averaged for each native contact over 500 simulation runs for each system. The number assigned to the NC pair code (*x* axis) corresponds to the native contact listed in Table 2. Native contacts in chain 1 (●), native contacts in chain 2 (▲), native contacts shared by chains 1 and 2 (■). Encircled symbols represent native contacts that that form the critical folding nuclei. Simulations were performed at $T = 0.25$.

different folding process, closer to the one observed for the unsplit protein, was observed for the Mid- and C-split cases. In these systems, both chains coassembled from the beginning of the folding process and jointly proceeded to the folded state. The fact that folding for the N-split system was significantly inhibited (relative to the parent 48-mer protein and to the other two split protein systems) further supports the notion that folding is less efficient when individual folding of one of the fragments (i.e., the nuclei-containing fragment) occurs. The mechanistic insight obtained by this analysis is consistent with our interpretation of the folding landscapes and snapshots shown in Fig. 6.

It is worth noting that in all the split protein cases, contact 58 (where each protein is split; see Table 1) was one of the very last native contacts to form in the folding process, as reflected by the very short τ_f associated with its formation (Fig. 7). Interestingly, all other native contacts that were locally affected upon protein fragmentation in each system also formed at relatively short τ_f, toward the very end of the folding process; these contacts included pair codes 9, 23, 28, and 29, pair codes 45 and 37, and pair code 50 for the N-, Mid-, and C-split systems, respectively (Table 1). Additionally, although contact 28 was one of the last to form in the N-split system, this contact was the first to form in both the Mid- and C-split systems. Most noteworthy are the observations that reattachment at (or near) the split site occurred late in all the split folding processes and that formation of interchain nuclei contacts occurred early in the cases of productive folding (i.e., the Mid- and C-split cases). This confirmed that efficient folding depends on the early ''gluing'' of the fragments specifically by the early interchain formation of folding nuclei contacts. Furthermore, productive folding appears to be independent of the early reconstitution of the original full-length 48-mer sequence, by reattachment of the fragments at the site where they were split.

Importance of folding nuclei in fragment reassembly of a split 64-mer

To test whether a shared folding nucleus contributed to the reassembly efficiency of proteins other than the 48-mer, we analyzed a model 64-mer (41,43,44). It is important to note that, like the 48-mer, this 64-mer also folds according to a classical nucleation mechanism where the core of critical native contacts that forms at an early stage of the folding process is composed of residues 2, 3, and 24–37, which have >90% probability of forming native contacts in the transition-state intermediates (34). Also noteworthy is that in contrast to the folding nucleus of the 48-mer, the amino acid composition of the folding core of the 64-mer is only 50% hydrophobic, and its location is on the side (as opposed to the center) of the folded structure. Additionally, given the larger size of this sequence relative to the 48-mer, it exhibits a more complex and therefore slower pattern of folding, where 81 native contacts characterize the folded structure.

To evaluate the importance of the folding nucleus in the reconstitution of a split 64-mer, two symmetric two-fragment systems, each containing a 27-mer and a 37-mer fragment, were derived (Fig. 8). N-split$_{64}$ was derived by splitting the 64-mer near the N-terminus of the sequence between residues 27 and 28, whereas C-split$_{64}$ was derived by splitting the sequence toward the C-terminal end of the sequence between residues 37 and 38. The additional native contact that restores the amino acid connection lost upon excision in each fragmentation case changes the native energy corresponding to the parent 64-mer from $-30.13k_B$T to $-29.93k_B$T and $-30.22k_B$T for the N-split$_{64}$ and C-split$_{64}$ systems, respectively. It is important to note that all of the 13 native core contacts of C-split$_{64}$ form within the larger of the two chains, whereas 8 out of 13 core contacts (>60%) of N-split$_{64}$ form between the two fragments, and only 5 out of 13 core contacts form within a single chain (two contacts in the shorter chain and three contacts in the longer chain (see Supplementary Material, Table 1S).

Given the distribution of core contacts for the N- and C-split$_{64}$, we hypothesized that folding would be more efficient in the case of the N-split$_{64}$ due to the higher number of interchain critical native contacts in this system relative to the C-split$_{64}$. Indeed, the resulting MFT, calculated as the average over 100 simulations at $T = 0.22$ (a temperature

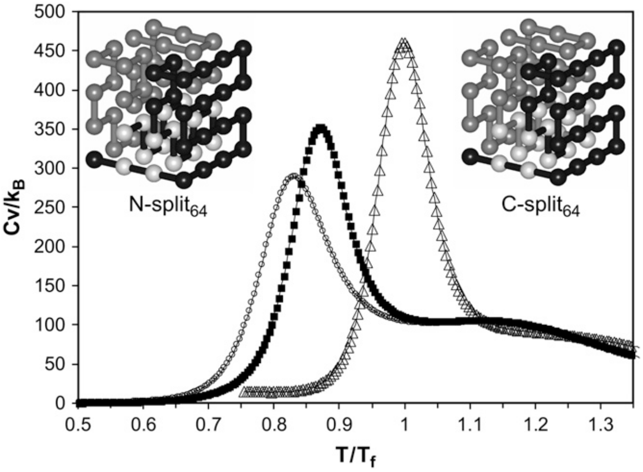

FIGURE 8 Thermodynamic analysis of 64-mer systems. Heat capacities for the 64-mer (△), N-split$_{64}$ (■), and C-split$_{64}$ (○) proteins as a function of temperature. The sequence of this protein is: KEKSTAGRVASGVLDSVACGVLGDIDTLQGSPIAKLKTFYGNKFNDVEASQAHMIR WPNYTLPE. Peaks represent transition temperatures (T_{max}) for each system. All temperatures are normalized by 0.27, the folding temperature of the unsplit 64-mer (T_f). Normalized transition temperatures were found to be $T_{max}/T_f = 1$ for the 64-mer, $T_{max}/T_f = 0.856$ for the N-split$_{64}$ system, and $T_{max}/T_f = 0.815$ for the C-split$_{64}$ system. Thermodynamic simulations were performed for 5E10 MC steps (a long time relative to folding times). (*Insets*) Schematic representations of the folded structures. Amino acids shown in black or connected by black lines correspond to those in the first fragment (chain 1); amino acids shown in dark gray or connected by dark gray lines correspond to those in the second fragment (chain 2); amino acids shown in light gray correspond to those that form the folding nucleus.

below the $T_{\max}$ for the two systems) for protein reassembly within 5×10^8 MC steps, was $3.10 \pm 0.48 \times 10^8$ for the N-split$_{64}$ and $3.99 \pm 0.82 \times 10^8$ for the C-split$_{64}$. Both split cases exhibited slower folding kinetics relative to that of the unsplit 64-mer (MFT = $1.38 \pm 0.08 \times 10^8$) at the same temperature. Moreover, the N-split$_{64}$ protein was observed to reassemble in 66 out of 100 simulation trials (FF = 66%) with an AFR of 2.13×10^{-7}, whereas the C-split$_{64}$ protein only reassembled 59 times out of 100 trials (FF = 59%) with an AFR of 1.48×10^{-7}, indicating a 31% decrease in folding for the C-split$_{64}$ relative to the N-split$_{64}$. It is also important to note that a decrease in thermal stability was observed upon fragmentation of the 64-mer, as reflected by the much lower $T_{\max}$ of the N-split$_{64}$ ($T_{\max} = 0.23$) and C-split$_{64}$ ($T_{\max} = 0.22$), relative to that of the unsplit 64-mer ($T_{\max} = 0.27$). Furthermore, a small and broad C_{v} peak is observed for the split systems, which implies an increase in near-native conformations. This effect suggests that their thermal transition is less cooperative (42). However, the split systems still follow a two-state mechanism, which is evidenced by the presence of a single C_{v} peak. Thus, the split 64-mer systems exhibited the same correlation between thermal stability and folding kinetics as was observed for the split 48-mer system.

CONCLUSIONS

In this work, we used two relatively simple model systems to obtain insight about how the choice of split sites affects the thermodynamics and kinetics of protein reassembly and folding upon fragmentation. Specifically, we focused our studies on understanding how the splitting of critical native contacts, which are located in the critical core that leads to folding, contribute to productive folding. In general, our results showed that the folding process for different split fragment systems is slower relative to the case of an unsplit protein, consistent with experimental observations (10,11,17). Furthermore, the nature and magnitude of reassembly retardation was highly dependent on the distribution of the critical nuclei between the two split fragments. Strategic splitting of the critical core was shown to 1), prevent the permanent preassembly of an individual fragment that would otherwise inhibit the assembly of the two chains; and 2), drive the formation of interchain native contacts that lead to productive folding. The importance of a shared folding core was particularly evident by the slower folding kinetics that were observed in the N-split system, where the critical core was localized in a single fragment, as compared with the C-split system, where the critical core was more equally shared between the two fragments.

Although a precise characterization of the folding mechanism or of the transition states for the N-, Mid-, and C- split systems was not determined, we observed that the concentration of the core native contacts in a single fragment changed the folding mechanism from a cooperative coassembly process, where the two fragments fold together, to a two-step assembly process, where an individual chain preassembles and then forms interchain connections with the second chain. Coassembly was observed for the fastest folding (C-split and Mid-split) systems, whereas a two-step folding mechanism was observed for the slowest folding (N-split) protein. Although these results raise the possibility that coassembly is more effective than a two-step assembly process for this model 48-mer, what appears to be most important is that the coassembly mechanism of the C- and Mid-split proteins deviates the least from the efficient folding mechanism of the unsplit parent protein, where the entire sequence folds together given that all its amino acids are connected. In other words, the sharing of the core between the two fragments contributes to preserving the overall folding mechanism exhibited by the parent protein so that the process is still productive when the protein is fragmented. In the future, it would be interesting to determine the exact mechanism by which folding occurs in the split protein systems by calculating committor probabilities, which quantify the tendency of a configuration along the path to relax to the native state under the systems' intrinsic dynamics (30,45,46). This type of analysis would lead to more detailed understanding of how the reassembly process deviates from the case of unsplit protein folding, even in those cases where the folding kinetics are only minimally affected. Additionally, it would be instructive to study other characteristics of multichain protein assembly that are still poorly understood, such as the way in which the order of events (i.e., native contact formation, which chains come together first, etc.) and the concentration of each chain affect the reassembly process.

The results obtained for the two split cases of the 64-mer shed additional light on the importance of distributing the folding nuclei between the split domains when fragmenting proteins. The fact that the same kinetic and thermodynamic trends that were observed in the 48-mer were also observed for the 64-mer split systems suggests that the enhancement in protein reassembly that results from an interchain distribution of the folding nucleus is not unique to this particular 48-mer and might apply generally to proteins that follow a classical nucleation folding mechanism. Interestingly, the role of the folding core in fragment reassembly was highly relevant in both of the systems we tested, despite the fact that in the case of the 48-mer the nucleus was highly hydrophobic and located near the center of the folded structure and in the case of the 64-mer the nucleus was highly hydrophilic and found off-center of the folded structure.

From the standpoint of application, this analysis suggests that protein reassembly could be made more efficient by fragmenting proteins in regions where the critical folding nucleus could be distributed in such a way as to naturally ''glue'' the fragments during the folding process. The dependency on the reconstitution of the core for folding observed in our simulations parallels the ''folding by binding'' mechanism that has been shown for tandem homodimeric proteins that share large interfaces upon binding (47,48). In the case of these protein complexes, often referred to as obligatory dimers, the association and formation of a large interface between two monomers (that remain largely unfolded by themselves) are key prerequisites for the concurrent folding of the two chains as one stable complex. The importance of driving fragments together to form productive interchain interactions by strategic splitting of the core is supported experimentally by (and could be considered somewhat analogous to) the assisted reassembly of several split proteins like GFP, DHFR, and ubiquitin by the addition of leucine zippers to each fragment that serve to enhance fragment interactions. It is interesting to note that the reassembly of the split ubiquitin protein is observed experimentally when the protein is fragmented such that ~60% of the amino acid residues that make up the compact hydrophobic core (analogous to the folding nucleus) are located in one fragment and 40% in the other fragment (12). The folding core of ubiquitin has been previously identified experimentally and computationally (49,50). Moreover, since the process of identifying the folding nuclei of a real protein experimentally remains a challenge, it is interesting to speculate that amino acid interactions that are key to the folding process (i.e., a folding core) of a target protein can be identified by correlating changes in reassembly kinetics (which are relatively easy to measure) to different splitting patterns (made by fragmenting the protein at different sites). This type of analysis would be analogous to the process of identifying the catalytic site in an enzyme by systematic amino acid mutation and evaluation of the functionality of the protein by activity assays.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

The authors thank Samuel Edoho-Eket for his assistance in code writing and analysis.

L.M.C. is grateful to the Jane and John C. Coleman Fellowship Program and the National Science Foundation Integrative Graduate Education and Research Traineeship Fellowship Program. E.E.B. is grateful to the Sloan Foundation for a Graduate Fellowship and the Department of Education for a Graduate Assistance in Areas of National Need Fellowship. This work was also supported by National Science Foundation awards BES-0093769 and ECS-0304483 (to F.E.) and by a Beckman Young Investigator Award (to M.P.D.).

REFERENCES

1. Fields, S., and O. Song. 1989. A novel genetic system to detect protein-protein interactions. *Nature*. 340:245–246.
2. Rual, J. F., K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*. 437:1173–1178.
3. Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. 2003. A protein interaction map of *Drosophila melanogaster*. *Science*. 302: 1727–1736.
4. Li, S., C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science*. 303:540–543.
5. Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*. 98:4569–4574.
6. Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*. 403:623–627.
7. McCraith, S., T. Holtzman, B. Moss, and S. Fields. 2000. Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA*. 97:4879–4884.
8. Bartel, P. L., J. A. Roecklein, D. SenGupta, and S. Fields. 1996. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet*. 12:72–77.
9. Kerppola, T. K. 2006. Complementary methods for studies of protein interactions in living cells. *Nat. Methods*. 3:969–971.
10. Ghosh, I., A. D. Hamilton, and L. Regan. 2000. Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *J. Am. Chem. Soc*. 122:5658–5659.
11. Magliery, T. J., C. G. Wilson, W. Pan, D. Mishler, I. Ghosh, A. D. Hamilton, and L. Regan. 2005. Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: scope and mechanism. *J. Am. Chem. Soc*. 127:146–157.
12. Johnsson, N., and A. Varshavsky. 1994. Split ubiquitin as a sensor of protein interactions *in vivo*. *Proc. Natl. Acad. Sci. USA*. 91:10340–10344.
13. Pelletier, J. N., F. X. Campbell-Valois, and S. W. Michnick. 1998. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc. Natl. Acad. Sci. USA*. 95:12141–12146.
14. Galarneau, A., M. Primeau, L. E. Trudeau, and S. W. Michnick. 2002. β-lactamase protein fragment complementation assays as *in vivo* and *in vitro* sensors of protein protein interactions. *Nat. Biotechnol*. 20:619–622.
15. Wehrman, T., B. Kleaveland, J. H. Her, R. F. Balint, and H. M. Blau. 2002. Protein-protein interactions monitored in mammalian cells via complementation of β-lactamase enzyme fragments. *Proc. Natl. Acad. Sci. USA*. 99:3469–3474.
16. Paulmurugan, R., and S. S. Gambhir. 2003. Monitoring protein-protein interactions using split synthetic renilla luciferase protein-fragment assisted complementation. *Anal. Chem*. 75:1584–1589.

17. Deo, S. K. 2004. Exploring bioanalytical applications of assisted protein reassembly. *Anal. Bioanal. Chem.* 379:383–390.

18. Kerppola, T. K. 2006. Visualization of molecular interactions by fluorescence complementation. *Nat. Rev. Mol. Cell Biol.* 7:449–456.

19. Wilson, C. G., T. J. Magliery, and L. Regan. 2004. Detecting protein-protein interactions with GFP-fragment reassembly. *Nat. Methods.* 1: 255–262.

20. Hennecke, J., P. Sebbel, and R. Glockshuber. 1999. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.* 286:1197–1215.

21. Iwakura, M., T. Nakamura, C. Yamane, and K. Maki. 2000. Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.* 7:580–585.

22. Betton, J. M., and M. Hofnung. 1994. *In vivo* assembly of active maltose binding protein from independently exported protein fragments. *EMBO J.* 13:1226–1234.

23. Paszkiewicz, K. H., M. J. Sternberg, and M. Lappe. 2006. Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics.* 22:1353–1358.

24. Cabantous, S., T. C. Terwilliger, and G. S. Waldo. 2005. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* 23:102–107.

25. Paulmurugan, R., and S. S. Gambhir. 2007. Combinatorial library screening for developing an improved split-firefly luciferase fragment-assisted complementation system for studying protein-protein interactions. *Anal. Chem.* 79:2346–2353.

26. Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry.* 33:10026–10036.

27. 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.

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

29. Neira, J. L., B. Davis, A. G. Ladurner, A. M. Buckle, P. Gay Gde, and A. R. Fersht. 1996. Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. Direct evidence for a nucleation-condensation mechanism. *Fold. Des.* 1:189–208.

30. Borrero, E. E., and F. A. Escobedo. 2006. Folding kinetics of a lattice protein via a forward flux sampling approach. *J. Chem. Phys.* 125: 164904.

31. Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1996. Improved design of stable and fast-folding model proteins. *Fold. Des.* 1:221–230.

32. Bagci, Z., R. L. Jernigan, and I. Bahar. 2002. Residue coordination in proteins conforms to the closest packing of spheres. *Polymer (Guildf.).* 43:451–459.

33. Contreras Martinez, L. M., F. J. Martinez-Veracoechea, P. Pohkarel, A. D. Stroock, F. A. Escobedo, and M. P. DeLisa. 2006. Protein translocation through a tunnel induces changes in folding kinetics: a lattice model study. *Biotechnol. Bioeng.* 94:105–117.

34. Borrero, E. E., and F. A. Escobedo. 2007. Reaction coordinates and transition pathways of rare events via forward flux sampling. *J. Chem. Phys.* 127:164101.

35. Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 18:534–552.

36. Frenkel, D., and B. Smit. 2001. Understanding Molecular Simulation: from Algorithms to Applications. Academic Press, San Diego.

37. Bratko, D., and H. W. Blanch. 2001. Competition between protein folding and aggregation: a three-dimensional lattice-model simulation. *J. Chem. Phys.* 114:561–569.

38. Sali, A., E. Shakhnovich, and M. Karplus. 1994. How does a protein fold? *Nature.* 369:248–251.

39. Earl, D. J., and M. W. Deem. 2005. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7:3910–3916.

40. Ferrenberg, A. M., and R. H. Swendsen. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63:1195–1198.

41. Cellmer, T., D. Bratko, J. M. Prausnitz, and H. Blanch. 2005. Protein-folding landscapes in multichain systems. *Proc. Natl. Acad. Sci. USA.* 102:11692–11697.

42. Kaya, H., and H. S. Chan. 2000. Energetic components of cooperative protein folding. *Phys. Rev. Lett.* 85:4823–4826.

43. Cellmer, T., D. Bratko, J. M. Prausnitz, and H. Blanch. 2005. Thermodynamics of folding and association of lattice-model proteins. *J. Chem. Phys.* 122:174908.

44. Leonhard, K., J. M. Prausnitz, and C. J. Radke. 2003. Solvent-amino acid interaction energies in 3D-lattice MC simulations of model proteins. Aggregation thermodynamics and kinetics. *Phys. Chem. Chem. Phys.* 5:5291–5299.

45. Li, L., and E. I. Shakhnovich. 2001. Different circular permutations produced different folding nuclei in proteins: a computational study. *J. Mol. Biol.* 306:121–132.

46. Du, R., V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–350.

47. Levy, Y., A. Caflisch, J. N. Onuchic, and P. G. Wolynes. 2004. The folding and dimerization of HIV-1 protease: evidence for a stable monomer from simulations. *J. Mol. Biol.* 340:67–79.

48. Levy, Y., P. G. Wolynes, and J. N. Onuchic. 2004. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA.* 101: 511–516.

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

50. Michnick, S. W., and E. Shakhnovich. 1998. A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* 3:239–251.