Intermediates and the folding of proteins L and G

SCOTT BROWN¹ AND TERESA HEAD-GORDON

Department of Bioengineering, University of California (UC), Berkeley, Berkeley, California 94720-1762, USA

(RECEIVED July 17, 2003; FINAL REVISION January 12, 2004; ACCEPTED January 13, 2004)

Abstract

We use a minimalist protein model, in combination with a sequence design strategy, to determine differences in primary structure for proteins L and G, which are responsible for the two proteins folding through distinctly different folding mechanisms. We find that the folding of proteins L and G are consistent with a nucleation-condensation mechanism, each of which is described as helix-assisted β -1 and β -2 hairpin formation, respectively. We determine that the model for protein G exhibits an early intermediate that precedes the rate-limiting barrier of folding, and which draws together misaligned secondary structure elements that are stabilized by hydrophobic core contacts involving the third β -strand, and presages the later transition state in which the correct strand alignment of these same secondary structure elements is restored. Finally, the validity of the targeted intermediate ensemble for protein G was analyzed by fitting the kinetic data to a two-step first-order reversible reaction, proving that protein G folding involves an on-pathway early intermediate, and should be populated and therefore observable by experiment.

Keywords: intermediates; kinetic mechanism; protein L and G, minimalist model; protein folding

Although thermodynamics and kinetics of small proteins that fold via a two-state manner are reasonably well understood (Gruebele 2002b; Myers and Oas 2002; Daggett and Fersht 2003a), understanding how (and why!) proteins fold through intermediates will be especially relevant for larger proteins, more complicated topologies, and their possible connection to aggregation processes that are responsible for disease (Speed et al. 1997). Some of the open questions surrounding intermediates include the detection of the so-called “hidden” intermediates by kinetic experiments, whether intermediates can occur earlier than the rate-limit-

ing step in folding, that is, do free energy barriers that precede the rate-limiting nucleation barrier of the folding reaction exist, and if they are “off-pathway,” and therefore obstruct the functionally important progress of folding (Gruebele 2002a; Ozkan et al. 2002; Qin et al. 2002; Sanchez and Kiefhaber 2003a).

This work examines the question of intermediates by simulating the folding of two members of the ubiquitin fold class, Ig-binding proteins L and G. Proteins L and G make excellent targets for theoretical study as their folding attributes have been extensively studied by experiment (Gu et al. 1995, 1997; Park et al. 1997, 1999; Scalley et al. 1997; Kim et al. 2000; McCallister et al. 2000; Krantz et al. 2002). These two single-domain proteins have little sequence identity but identical fold topologies, consisting of a central α -helix packed against a four-strand β -sheet composed of two β -hairpins. Experimental evidence indicates that protein L folds in a two-state manner through a transition state ensemble involving a native-like β -hairpin 1, and largely

---

Reprint requests to: Teresa Head-Gordon, Department of Bioengineering, 472 Donner Laboratory, University of California, Berkeley, Berkeley, CA 94720-1762, USA; e-mail: TLHead-Gordon@lbl.gov; fax: (510) 486-6632.

¹Present address: Abbott Laboratories, 1401 Sheridan Road, North Chicago, IL 60064-4000, USA.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.03316004>.

disrupted β -hairpin 2 (Gu et al. 1997; Scalley et al. 1997; Kim et al. 2000). Protein G, on the other hand, folds through a possible early intermediate (Park et al. 1997, 1999; Speed et al. 1997), followed by a rate-limiting step that involves formation of β -hairpin 2. They therefore provide a perfect contrast to understand features that give rise to protein folding intermediates, while controlling for size and topology.

There have been a number of recent simulations of coarse-grained models of proteins L and/or G using different forms of minimalist models (Head-Gordon and Brown 2003). Shimada and Shakhnovich (2002) have used ensemble dynamics to characterize the kinetics of protein G using an all-atom G $\ddot{o}$ potential. Karanicolas and Brooks (2002) use a G $\ddot{o}$ potential bead model supplemented with sequence-dependent MJ statistical potentials to differentiate the folding of G and L. They found the origin of asymmetry in the folding of protein L and G to be in concurrence with that found by Nauli et al. (2001), who used a computer-based design strategy to reengineer the protein G sequence to include more stabilizing interactions for the first β -hairpin turn, producing a protein more faithful to the mechanism of folding for protein L.

Our recent work, inspired by early efforts of Thirumalai and co-workers (Honeycutt and Thirumalai 1990; Guo et al. 1992; Guo and Thirumalai 1996), develops physics-based potentials that make the connection between free energy landscapes and amino acid sequence, allowing us to engineer sequences that fold into α -helical, β -sheet, and mixed α/β protein topologies (Sorenson and Head-Gordon 1999, 2000, 2002a,b; Brown et al. 2003). These coarse-grained protein models provide the right emphasis of the most relevant native state features (Head-Gordon and Brown 2003) by capturing the correct spatial distribution of local and nonlocal contacts that are considered to be possibly the most important in governing the overall kinetics of protein folding (Plaxco et al. 1998; Alm et al. 2002). We have previously explored its use for members of the ubiquitin α/β fold class including proteins L and G and ubiquitin (Sorenson and Head-Gordon 2000, 2002a,b; Brown et al. 2003). Recently, we have verified that the use of a three-letter sequence code is capable of translating the differences in primary sequence for proteins L and G into the experimentally observed differences in thermodynamic and kinetic properties of folding (Brown et al. 2003).

In this work, we analyze the kinetics of folding for these distinct sequences to characterize the dynamics of navigating the free-energy landscape from unfolded to native state. P_{fold} simulations (Du et al. 1998) and contact map analysis have been used to characterize the folding landscape. The folding of our protein L model follows two-state kinetics, and shows the presence of a transition-state ensemble with a well-formed β -hairpin 1. Similar analysis of protein G shows that it folds through at least two pathways, which we label the fast and slow pathways. The fast pathway exhibits

two-state kinetics, and folds through a transition-state ensemble with a well-formed β -hairpin 2. In both cases of protein L and fast-folding protein G, these secondary structural elements are assisted by the α -helix, and the overall folding mechanism seems consistent with a nucleation-condensation mechanism observed for other proteins (Daggett and Fersht 2003a,b).

The slow pathway for protein G is what gives rise to three-state kinetics, and involves an early intermediate, that is, an intermediate that precedes the rate-limiting step in folding. The characteristics of the intermediate are hydrophobic contacts involving the third β -strand interacting with β -strands 1 and 2, although the associated secondary structure strand elements are misaligned relative to our model of the folded state. The transition state that occurs after the intermediate and proceeding the folding to the native state is characterized by native-like registering of these same β -strand pairings. The tractability of the simulation model allows us to fit the kinetic data to a unimolecular two-step kinetic model to summarize the kinetics of protein G folding. We confirm that a barrier, in fact, separates the unfolded state from the early folding intermediate, and is lower in free energy relative to the unfolded state, and that the intermediate should be populated and observable by experiment.

Results and Discussion

One of the differences between our L and G model proteins is manifested in the relative thermodynamic stability of the different elements of secondary structure. Figure 1 shows free-energy projections along $\chi_{\beta-1}$ and $\chi_{\beta-2}$ for L and G. From these projections there is a minimum free-energy path connecting the unfolded ensemble to the folded ensemble that involves either sequential formation of β -hairpin 1 followed by β -hairpin 2 (protein L), or β -hairpin 2 followed by β -hairpin 1 (protein G). However, we appear to only be getting part of the picture in Figure 1, as the barrier height separating the unfolded and folded ensembles is insufficiently high (relative to $k_B T$) to justify locating the rate-limiting transition state solely on this surface.

The folding kinetics at the folding temperature, shown in Figure 2, illustrates the difference in folding mechanism between L and G (fit parameters given in Table 2). The kinetic data for protein L are fit well by a single exponential, consistent with what is reported in the literature for protein L (Scalley et al. 1997; Kim et al. 2000). Thus, our protein L model folds in a cooperative two-state manner, and possibly through the initial formation of β -hairpin 1.

For protein G, the story is not as straightforward. We find that protein G folds slower than protein L by a factor of 2, qualitatively consistent with the experiment (McCallister

Figure 1. Free-energy projections onto order parameters $\chi_{\beta-1}$ and $\chi_{\beta-2}$ for L and G. (A) Free-energy contour plot for protein L as a function of native-state similarity of the second (C-terminal) β -sheet region $\chi_{\beta-2}$ and first (N-terminal) β -sheet region $\chi_{\beta-1}$ at the folding temperature. Note the minimum free-energy path connecting the unfolded and folded ensembles proceeds through a transition state in which the β -1 region is native and the β -2 region is largely disrupted. (B) Free-energy contour plot for protein G as a function of native-state similarity of the second (C-terminal) β -sheet region $\chi_{\beta-2}$ and first (N-terminal) β -sheet region $\chi_{\beta-1}$ at the folding temperature. For G, the minimum free-energy path connecting the unfolded and folded ensembles proceeds through a transition state in which the β -2 region is native-like and the β -1 region is disrupted. Contour lines are spaced $k_B T$ apart.

et al. 2000); however, the kinetic data for protein G are better fit by at least a double exponential (parameters shown in Table 2). From this fit we find two populations; one involves a fast-folding event in which roughly 80% of the population folds cooperatively, and a slow-folding remainder of the ensemble that folds by a different mechanism that we analyze further below. The time scale that serves to roughly delineate these two populations is 2×10^6 time steps. After this many time steps the majority of the fast folding states have folded, while only a tiny fraction of the slow folders have folded. Using 2×10^6 time steps as a

Table 1. Parameters obtained from fits to kinetic data

	T	A_0	$1-A_0$	τ_0	τ_1	$\chi^2/10^{-4}$
L	0.42	1.0	0	15,700	0	3.43
G	0.41	0.81	0.19	13,700	46,400	0.353

The data is fit to the equation: $A_0 \exp(-t/\tau_0) + (1 - A_0) \exp(-t/\tau_1)$.

cutoff, we refit the kinetic data of the fast-pathway population for protein G and obtain a single exponential, while the fit to the remaining 20% of slow folders gives a double exponential, suggestive of an intermediate state in the slow folding trajectories.

Figure 2. Folding kinetics for proteins L and G. (A) Fraction of folded states P_{fold} as a function of time t for protein L at the folding temperature. The best fit to the data is by a single exponential. (B) Fraction of folded states P_{fold} as a function of time t for protein G at its folding temperature. The best fit for this data is to a double exponential. All fit parameters are given in Table 2.

Table 2. Order parameters, Q , used for characterizing folding mechanisms in proteins *L* and *G*, along with contacts used to define them

Q	i th: j th Bead contacts
Protein L transition state	
$\beta_1\beta_2\alpha$	6:16, 6:17 6:18 7:15 7:16 7:17 8:13 8:14 8:15 8:16 8:17 9:1 9:14 9:15 10:14 10:15 20:24 23:27 29:33 30:34
Protein G transition state (fast pathway)	
$\beta_3\beta_4\alpha$	10:14 20:24 20:27 23:27 24:28 27:31 36:53 36:54 36:55 37:53 38:52 38:53 39:51 41:48 41:50 41:51 42:48 42:49 43:47 43:48 43:49 44:48
Protein G transition state (slow pathway)	
$\beta_1\beta_3\beta_2$	8:14 8:15 8:16 8:17 8:36 8:37 8:38 9:13 9:14 9:15 9:36 10:14 14:36 15:36 17:38 19:40 19:41 19:42 20:41 20:42 20:49 21:41 21:42 21:43 43:48
Protein G intermediate	
$\beta_2\beta_3\alpha$	8:14 9:13 9:14 9:15 10:14 18:40 18:41 19:39 19:40 19:41 23:27 27:31 31:35 42:48 43:47 43:48 43:49 44:48

If we examine the folding ensemble at both early and late times for the two populations, we see that the fast pathway involves a collapse concomitant with folding scenario (Fig. 3a). The fast pathway also involves a greater degree of

native α -helix formation relative to the slower population. This is an important difference between the two pathways, and we will return to it later. The slower population is characterized by more relative ordering of both β -sheet re-

Figure 3. Shows the presence of two folding pathways for protein G. The fast pathway corresponds to a collapse concomitant with folding scenario, while the slow pathway corresponds to nonproductive collapse and a longer process of finding the native structure.

Figure 4. Free-energy surface projected onto order parameters R_g and χ . (A) Free-energy contour plot for protein L as a function of radius of gyration R_g and native-state similarity χ . In this plot there is only a single dominant minimum that corresponds to a collapsed, largely native structure. (B) Free-energy contour plot for protein G as a function of radius of gyration R_g and native-state similarity χ . In this plot there appear to be two dominant minima, one corresponding to collapsed nonnative structures and the other to collapsed native-like structures.

gions 1 and 2 (Fig. 3B). Note that both kinetic pathways exhibit a more developed β -2 region relative to the β -1 region, reflecting what is seen in the thermodynamic analysis. The picture from both kinetics and thermodynamics appears to be consistent, and points to a folding mechanism that involves formation of β -hairpin 2 at some rate-limiting step prior to that of β -hairpin 1.

Shown in Figure 4 is a two-dimensional free-energy surface projected onto the radius of gyration, R_g , and native-state similarity parameter, χ . Figure 4A shows the relationship between L collapse and native-state formation, which appears to occur by a single pathway leading from expanded, nonnative to the minimum on the surface corre-

sponding to collapsed and native-like. This is consistent with the picture from kinetic data of a collapse concomitant with folding scenario (Plaxco et al. 1999). In contrast to this, for protein G there appear to be two pathways for collapse, with two separate minima for each pathway, as illustrated by the arrows in Figure 4B. One pathway involves collapse to a largely nonnative structure, whereas the other pathway reflects a collapse concomitant with folding scenario, as seen with protein L. The barrier separating these two minima in Figure 4B again has insufficient height to account for observed kinetic data.

Recent work has strongly emphasized that the choice of reaction coordinate for monitoring folding progress is important for the observation of intermediates and general interpretation of kinetic data (Qin et al. 2002). A potential pitfall of choosing a reaction coordinate is illustrated in Figure 5, which shows the potential of mean force for protein G as a function of native-state similarity in going from the unfolded ($\chi \approx 0$) to folded ($\chi \approx 1$) states for a range of temperatures spanning the folding temperature. In producing our kinetic data we use this same native-state similarity parameter to determine the extent of folding during folding trajectories. Note that the folding temperature for the protein G sequence is $T^* \approx 0.41$. Jumping from the free-energy surface at $T^* = 0.5$ to the surface at $T^* = 0.35$ would involve a downhill rearrangement in the distribution of the unfolded ensemble. These results represent an alternative interpreta-

Figure 5. Potential of mean force vs. native state similarity as a function of temperature for protein G. The folding temperature is $T^* = 0.41$. Based on this projection we might conclude that there is a shift in the unfolded population as we approach folding conditions. There is also evidence for a small barrier.

tion of ultrafast folding experiments that remains consistent with overall evidence for two-state folding (Parker and Marqusee 1999).

One of the benefits of coarse-grained models is the ability to fully characterize ensemble kinetics on the free energy landscape by investigating transition state ensembles, and putative intermediates, provided we can find suitable reaction coordinates for their description. We examined a number of reaction coordinates before determining ones that adequately capture the folding events in our model. These include contact order parameters α , β -1, β -2, β -2 α , β -3 α , β -1 β -2, β -1 β -4, β -2 β -3, β -3 β -4, β -3 β -4 α , β -2 β -3 α , β -1 β -2 β -3, as well as a “diffuse” order parameter that is an expanded native state.

Several additional folding-trajectory analyses were performed to obtain more extensive kinetic characterization, in which progress during folding was monitored along a variety of these chosen order parameters. Structures with desired values along these order parameters were saved and served to form a set of putative transition states, which were then subsequently used as starting structures in trajectories for P_{fold} analysis. From the P_{fold} simulations we obtained a subset of successful order parameters (shown in Table 2), which correlate well with the definition of transition state ensemble. Through this procedure we determined a transition state ensemble for protein L, a transition state ensemble for the fast pathway in protein G, and the late transition state ensemble for the slow folding pathway of protein G. During P_{fold} simulations the trajectories either fold or do not fold, by definition. By saving the structures for those trajectories that did *not* fold in the P_{fold} simulations of structures corresponding to the transition state ensemble of the slow folding pathway for protein G, we were able to isolate the structural characteristics of the ensemble of early intermediates.

Figure 6A shows a contact map with reference lines indicating native-state contacts (black line) and the contacts that are present across at least 90% of the transition-state ensemble for protein L (gray line). The transition-state contacts show that the model for protein L folds through structures with a helix-assisted β -1 hairpin nucleus. Figure 6B shows a similar map of contacts to delineate the transition state ensemble in the fast pathway of protein G. In the case of protein G's fast pathway the transition-state en-

Figure 6. Contact map comparing native state (black) to contacts that are found present in the transition-state ensemble for 90% or greater of the structures (gray) for (A) protein L and (B) fast folding pathway of protein G.

semble involves formation of a helix-assisted β -2 hairpin nucleus.

Figure 7A shows the contact map for the contacts present in at least 90% of the structures in our intermediate ensemble. Figure 7B shows the contacts present in at least 90% of the late transition-state ensemble structures for the slow folding pathway of protein G. The intermediate is characterized by associated helix with β -strands 2 and 3, with a smaller amount of associated β -strands 1 and 3; however, these strands are misaligned relative to the native state. The subsequent transition state ensemble is in large part characterized by an alignment correction of this same strand association pattern exhibited in the intermediate, followed by more robust association of the other β -strands.

Finally, we prove that the intermediate occurs early on the pathway by fitting the data to a two-step reversible first-order UIN mechanism. Given the characterization of the intermediate for the slow pathway of protein G, we can

Table 3. Parameters obtained from fit to UIN kinetic model outlined in Appendix I to characterize the slow folding pathway of protein G

k_1	k_{-1}	k_2	k_{-2}	ΔG_{U-I}	ΔG_{I-N}
1.1×10^{-3}	1.3×10^{-4}	2.3×10^{-3}	4.0×10^{-6}	$-2.0 k_B T$	$-1.7 k_B T$

Figure 7. Contact map comparing native state (black) to contacts that are found present in the (A) intermediate ensemble and (B) transition-state ensemble, for 90% or greater of the structures for the slow folding pathway of protein G (gray).

monitor individual folding trajectories and record when states enter and leave the U, I, and N designations. Provided we observe a large number of trajectories, we can assemble a picture of the pathway the folding population follows as a function of time, and fit the corresponding data to the UIN mechanism:

The solutions for the time rate of change of concentration of each species expressed as a function of rate constants k_1 , k_{-1} , k_2 , and k_{-2} , is given in Appendix A. As far as we are aware, this is the first time a solution of the full UIN mechanism without simplifying approximations has appeared in the protein literature, although this mechanism is very often invoked in the analysis of protein folding reactions in its various simplified limiting forms. When our data is fit to these equations they yield values of the rate constants and associated estimates of relative free-energy minima, which are given in Table 4. We also show the quality of the fit of the slow-folding protein G data to the UIN model in Figure 8. Note that in Figure 8 we have enforced a restriction in which we eliminate all trajectories that fold prior to 2×10^6 time steps. This allows us to focus exclusively on the trajectories folding via the slow pathway, but leads to a slight anomaly in Figure 8 for the populations immediately prior to 2000τ . By removing greater than 95% of the fast folding trajectories we have excluded a small fraction of the slow-folding trajectories. In summary, the kinetic model demonstrably shows that a barrier, in fact, separates the unfolded state from the early-folding intermediate, and that the intermediate is lower in free energy relative to the unfolded state, and therefore should be populated and observable by experiment.

Conclusions

We find that protein L is a two-state folder, in agreement with existing experiments (Gu et al. 1997; Scalley et al. 1997; Kim et al. 2000). As such, it provides a unique reference system for understanding intermediates by comparing its folding to protein G, a structurally homologous protein of similar length for which continuous flow fluorescence experiments support the population of an early intermediate along the folding pathway (Park et al. 1997, 1999). This, by definition, involves the presence of an additional free energy barrier preceding the rate-limiting bar-

Table 4. Sequences for the minimalist models of protein L and G

Protein L						
1°	LBLBLBLBBN	NNBBBLBLBB	BNNNLLBLLB	BLLBNBLBLB	LBLNNNLBBL	BLBBBL
2°	EEEEETEHEH	THEEEEEEEE	HHEHHHHHHH	HHHEHTEEEE	EEETTTEEEE	EEEE
Protein G						
1°	LBLBLBLBBN	NNLBBLBLBB	BNNNLLBLLL	BLLBNBBLBL	BBBNNNLBBL	BLBLBL
2°	EEEEETEHEH	THEEEEEEEE	HHEHHHHHHH	HHHEHTEEEE	EEETTTEEEE	EEEE

Differences between the sequences are shown in bold.

Figure 8. Kinetic data and fits for the UIN folding mechanism scenario.

rier in folding. It is important to note that the stopped flow experiments cannot resolve any early intermediates in the folding of protein L, unlike the better time-resolved continuous flow experiments for protein G. Although continuous flow results have been called into question as a problem of suspect interpretation of ultrafast folding events in general (Krantz et al. 2002), our model supports the view that protein G folds through an early intermediate while protein L does not.

Protein L's transition state ensemble is composed of helix-assisted β -1 hairpin formation. We conclude that protein G folds through at least two pathways: a fast pathway involving roughly 80% of the folding population, with a transition state composed of a helix assisted β -2 hairpin nucleus, and a slow-folding pathway through which the remaining folding population proceeds in a three-state mechanism. Our model clearly demonstrates that the slow pathway involves the presence of an early intermediate involving the third β -strand, which is separated from the unfolded state by a significant barrier (relative to $k_B T$), and, in fact, is lower in free energy relative to the unfolded state (Table 4), and therefore, should be populated and observable by experiment. Therefore, our model strongly supports the interpretation of the continuous flow experiments by Park et al. (1997) as evidence of an early-folding intermediate.

Our results also emphasize that the choice of reaction coordinate used experimentally is very important to avoid conflicting conclusions concerning the presence of intermediates, as was found to be the case for reexamination of the presence of an intermediate in ubiquitin (Qin et al. 2002). Similar conclusions concerning the proper determination of reaction coordinates that monitor folding was also found by

Shimada and Shakhnovich (2002). Their simulation of protein G found that folding occurred through multiple pathways, each of which passes through an on-pathway intermediate. They showed that when folding is monitored by using burial of the lone tryptophan in protein G as the reaction coordinate, the ensemble kinetics shows a significant burst phase, while alternative reaction coordinates reveal the presence of different folding pathways. They make the point that ensemble averaging can mask the presence of multiple pathways when nonideal reaction coordinates are used. We required a variety of different order parameters, coupled with P_{fold} analysis, to characterize the reaction coordinates for protein L and protein G folding to find all intermediates and transition states. Furthermore, we fit our kinetic data to a UIN-type mechanism, and provide estimates of the rate constants and relative free-energy minima to fully characterize the folding pathways.

The work reported by Shimada and Shakhnovich (2002) using an all-atom Go potential most closely parallels the study described here of analyzing the folding of protein G. They observe three pathways, each involving its own intermediate: I_1 (helix-hairpin 1), I_2 (helix-hairpin 2), and I_3 (β -1- β -4), and that each pathway converges to the same transition state. Our physics-based α -carbon trace model finds two major pathways, each with its own transition state, with only one pathway exhibiting an intermediate characterized by β -2 β -3 α . At this point it is difficult to tell more about the structural nature of the experimental intermediate given the nonspecific nature of the tryptophan (on the third β -strand) reaction coordinate used in the experimental study. However, we expect that the structural details of the intermediate are potentially more reliably predicted with the all-atom simulation because our coarse-grained model inadequately describes β -sheet structure, and instead forms a β -strand bundle for proteins L and G.

However, due to the inexpensive cost of our bead model, we are able to perform various analyses of thermodynamics as well as P_{fold} analyses along entire trajectories to isolate structures belonging to the transition-state and intermediate ensembles. This level of detailed investigation is not possible (or is not pursued) in more complicated models. For example, the total number of trajectories examined in Shimada and Shakhnovich (2002) is only 50, whereas we examine 1000 folding trajectories. With only 50 trajectories we found that we were unable to reliably analyze our data and make comment on ensemble folding properties. We found this to be a particularly significant problem when fitting to our postulated three-state reaction mechanism. The primary advantage of physics-based bead models is that the kinetics and thermodynamics are fully characterizable with high-quality statistics, and the overall qualitative agreement with experiment is very good.

The differences in the folding properties of L and G, for the fast pathways, are consistent with a nucleation-conden-

sation model (Abkevich et al. 1994) or nucleation-collapse mechanism (Guo and Thirumalai 1995) that has been used to analyze kinetic data on two-state folders (Fersht 1997; Daggett and Fersht 2003a,b; Sanchez and Kiefhaber 2003b). Whereas the fast pathway mechanisms for L and G involve the contact-assisted formation of secondary structure to create a folding nucleus at the transition state, the slow pathway in protein G involves an obligatory intermediate that precedes the rate-limiting step, a result that may seem inconsistent with a nucleation-based mechanism. However, as is seen in the case of barnase (Daggett and Fersht 2003a), the intermediate for protein G assists in formation of the folding nucleus. It has been pointed out that increasing the hydrophobicity may lead to a shift in folding mechanism towards a molten globule-like intermediate (Daggett and Fersht 2003a); that does not appear to be the case here. It should be noted that in this model the sequences for proteins L and G have an identical number of L and B beads, and thus have an identical global hydrophobicity. However, the third β -strand is significantly more hydrophobic in protein G relative to protein L, and hence, the intermediate certainly arises due to stabilization by hydrophobic contacts.

This greater hydrophobicity for protein G helps stabilize an intermediate that draws together the secondary structure elements of β -strand 3 in association with β -strands 1 and 2, although these secondary structure elements are out of register relative to the native state. However, this helps set up the final step in folding, which now involves a transition-state ensemble that corrects for the misalignment of this core nucleus of associated strand elements. Recent work has suggested that intermediates that are higher in free energy relative to the unfolded state (perhaps hidden from experimental view) can accelerate folding (Wagner and Kiefhaber 1999; Sanchez and Kiefhaber 2003a). Protein G folding involves an intermediate that is more stable than the unfolded state, and in fact, slows down folding relative to protein L, all of which is supported by experiment as well as the coarse-grained model examined here. Perhaps hydrophobic-stabilized intermediates are a concession to certain amino acid sequences, designed by nature for other functional reasons, that would otherwise fold by enthalpic barriers that are simply too high.

Materials and methods

The protein model has been described in (Sorenson and Head-Gordon 1999, 2000, 2002a,b). The protein chain is modeled as a sequence of beads of three flavors—hydrophilic, hydrophobic, and neutral—designated by L, B, and N, respectively. In general, the pair-wise interaction between beads is attractive for hydrophobic-hydrophobic (B-B) interactions, and repulsive for all other bead pairs (although the strength of the repulsion interactions depends on the bead types involved). In addition to pair-wise nonbonded interactions, the other contributions to the potential energy func-

tion include bending and torsional degrees of freedom. The total potential energy function is given by

$$H = \sum_{\theta} \frac{1}{2} k_{\theta} (\theta - \theta_0)^2 + \sum_{\phi} \left[A(1 + \cos\phi) + B(1 - \cos\phi) + C(1 + \cos 3\phi) + D \left(1 + \cos \left[\phi + \frac{\pi}{4} \right] \right) \right] + \sum_{i,j \geq i+3} 4\epsilon_H S_1 \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - S_2 \left(\frac{\sigma}{r_{ij}} \right)^6 \right] \quad (2)$$

where ϵ_H determines the energy scale and sets the strength of the hydrophobic interactions. The bond angle energy term is a stiff harmonic potential with force constant $k_{\theta} = 20\epsilon_H/\text{rad}^2$, and $\theta_0 = 105^\circ$. The second term in the potential energy designates the torsional, or dihedral potential, and is given by one of the following: helical (H), with $A = 0$, $B = C = D = 1.2\epsilon_H$; extended (E), favoring β -strands, with $A = 0.9\epsilon_H$, $C = 1.2\epsilon_H$, $B = D = 0$; or turn potential (T), with $A = B = D = 0$, $C = 0.2\epsilon_H$. For each dihedral angle potential the global minimum is the specified secondary structure type, but has stable local minimum for the other secondary structure angles. This aspect of the potential sits between a Gō model and a purely ab initio energy function because the dihedral angle potential is assigned for each bead based on the known native state. However, we use no explicit secondary or tertiary structure template to define any aspect of the potential, and hence, the form and parameters are transferable to any protein. The nonbonded interactions are determined by $S_1 = S_2 = 1$ for B-B interactions; $S_1 = 1/3$ and $S_2 = -1$ for L-L and L-B interactions; and $S_1 = 1$ and $S_2 = 0$ for all N-L, N-B, and N-N interactions. For convenience all simulations are performed in reduced units, with mass m , length σ , energy ϵ_H , and k_B all set equal to unity. Note that although the nonbonded potential is symmetric with respect to inversion, that is, $V_{\text{nonbonded}}(r_{ij}) = V_{\text{nonbonded}}(r_{ji})$, this is not true for the dihedral interactions, as $\phi = f(r_i, r_{i+1}, r_{i+2}, r_{i+3})$. Thus, the total energy function is not symmetric with respect to indice permutations.

We perform constant-temperature simulations using Langevin dynamics in the low friction limit for characterizing the thermodynamics and kinetics of folding. Bond lengths are held rigid using the RATTLE algorithm (Andersen 1983). The free-energy landscape is characterized using the multiple, multidimensional weighted histogram analysis technique (Ferrenberg and Swendsen 1989; Kumar et al. 1995; Ferguson and Garrett 1999). We collect multidimensional histograms over a number of different order parameters, including energy E , radius of gyration R_g , and various native-state similarity parameters χ ,

$$\chi = \frac{1}{M} \sum_{i,j \geq i+4}^N h(\epsilon - |r_{ij} - r_{ij}^{\text{native}}|) \quad (3)$$

where the double sum is over beads on the chain, and r_{ij} and r_{ij}^{native} are the distances between beads i and j in the state of interest and the native state, respectively. h is the Heaviside step function, with $\epsilon = 0.2$ to account for thermal fluctuations away from the native state structure. M is a constant that satisfies the conditions that $\chi = 1$ when the chain is identical to the native state and $\chi = 0$ in the random coil state. The remaining χ parameters are specific to their respective elements of secondary structure. That is, χ_{α} involves summation over beads in the helix, and $\chi_{\beta-1}$ and $\chi_{\beta-2}$ involve summation over beads in the first β -sheet region and second β -sheet region, respectively, etc.

From the histogram method we get the density of states Ω , as a function of these order parameters, which can be used to calculate thermodynamic quantities. One quantity that is useful is the native-state population as a function of temperature

$$P_{Nat}(T) = \frac{\sum_{\substack{V, R_g, \chi \\ \chi_{\alpha} \chi_{\beta_1} \chi_{\beta_2}}} \Omega(V, R_g, \chi, \chi_{\alpha}, \chi_{\beta_1}, \chi_{\beta_2}) e^{-V/T}}{\sum_{\substack{V, R_g, \chi \\ \chi_{\alpha} \chi_{\beta_1} \chi_{\beta_2}}} \Omega(V, R_g, \chi, \chi_{\alpha}, \chi_{\beta_1}, \chi_{\beta_2}) e^{-V/T}} \quad (4)$$

where χ_{NBA} indicates the boundary of the native-state basin of attraction (NBA; Nymeyer et al. 1998). In constructing the free energy surfaces we collect histograms at 15 different temperatures: 1.20, 0.90, 0.70, 0.62, 0.60, 0.55, 0.50, 0.48, 0.46, 0.44, 0.42, 0.41, 0.40, 0.39, and 0.38. We run three independent trajectories at each temperature, and collect 10,000 data points per trajectory.

The kinetics of the folding process can be characterized by calculating a large number of first-passage times (the time required for a folding trajectory to first enter the native basin of attraction, defined to be $\chi_{NBA} = 0.40$). The first-passage times are calculated by taking an initial high-temperature random-coil structure and evolving it at the temperature of interest until recording the time that it first enters the native basin of attraction. We subtract off an initial correlation time in which the high-temperature chain is briefly equilibrated at the target temperature (this is the computational dead time during the kinetics run).

To accurately characterize the proper transition-state ensemble in our analysis of protein L and G folding, we employed the P_{fold} method proposed by Du et al. (1998). The method assigns a value, P_{fold} , to a particular structure corresponding to the probability that it will first fold to the native state before unfolding. Structures with P_{fold} values equal to 0.5 correspond to the transition-state ensemble for the model. To apply this method, we first sampled structures from our simulations corresponding to putative transition-state structures. "Putative" transition-state structures were originally isolated by requiring various combinations of order parameters to correspond to their maximum free-energy values in a one-dimensional projection of free energy against these order parameters. From this procedure, an ensemble of structures with $0.4 \leq P_{fold} \leq 0.6$ were isolated during multiple kinetic runs, and were defined as members of the transition state ensemble. By analyzing these structures we are able to postulate new reaction coordinates.

Identifying the transition state ensemble also allowed us to define an intermediate ensemble. By identifying those configurations that have $P_{fold} \approx 0.5$ we can save structures for the trajectories that fail to fold, thus allowing us to postulate an intermediate ensemble. The set of structures obtained in this way can be characterized by analyzing the contacts that are present across all members of the ensemble. Using the defining contacts we can test our definition of intermediates through the direct analysis of kinetic runs. The final test for the validity of any definition of an intermediate ensemble is an analysis performed by fitting to a two-step first-order reversible reaction, or UIN mechanism, $U \leftrightarrow I \leftrightarrow N$, where U is the unfolded state, I is the intermediate state, and N is the native state. The formal solution to this kinetic mechanism is given in Appendix A. The data to which we fit is obtained from simulation by monitoring the progress of the folding trajectories and marking

each time a state fits our definition of U, I, or N. We note that the kinetic analysis reported by Shimada and Shakhnovich fits the decay of the unfolded population separately from any of the I_1 , I_2 , or I_3 intermediate populations; that is, the fit violates mass balance.

Next, we discuss our sequence design procedure. Theoretical work (Bryngelson and Wolynes 1989; Sali et al. 1994; Onuchic et al. 1997) has elucidated a criterion for heteropolymers to be foldable by noting that there should be a significant energy gap between the native-state and average misfold energies. Our sequence design strategy makes use of this concept. We create a library of misfolds (obtained from simulation of multiple trajectories), and then maximize the energy gap, $\Delta E_{design} = \langle E_{misfold} \rangle - E_{native}$, through favorable mutations on the sequence. We start with a sequence that adopts the protein L/G target topology as given in Sorenson and Head-Gordon (2002b), and build upon it through sequence mutation to produce new sequences that comprise distinct members, protein L and G, within a target fold class (Brown et al. 2003). The sequence for protein L was determined by aligning it against the real protein L sequence (after mapping the 20-letter code to three-letter code as described in Brown et al. (2003), and proposing new mutations that moved the original sequence towards being more L-like. For protein G, all possible single mutations were investigated during the design process, with the final outcome resulting in the selection of mutations that were beads corresponding to errors in the protein G alignment (Brown et al. 2003). This is interesting in that it appears to hint at potential criteria for performing sequence mapping onto our minimalist code, which could allow for study of novel proteins whose structure is not yet known. Two of the five point mutations for L and G are shared in common (B18L and B47L), which serve to make the proteins more foldable and to clean up certain thermodynamic aspects of the original L/G sequence (Brown et al. 2003). Another three mutations are what serve to distinguish the sequence of protein L from that of protein G. Table 1 lists the sequences for L and G used in this study, in which there are a difference of six beads between the protein L and G sequences. The energy of the initial L/G sequence is $-32.4\epsilon_H$, while for the new protein L the native-state energy is $-28.8\epsilon_H$, and for protein G the native-state energy is $-26.9\epsilon_H$. For all native states we find that the energy distribution of the misfold library is well separated from the native-state energies.

Finally, we compare the structural similarity of the native state of our protein L model with the experimental structure. The root mean square distance (RMSD) between the native quenched structure of the protein L model and the protein L set of NMR solution structures (2PTL, residues 20–78) was found to be approximately 4.4 Å. This measure of RMSD was generated by the Combinatorial Extension (CE) Web server (<http://cl.sdsc.edu/ce.html>; Shindyalov and Bourne 1998). Calculating an RMSD between an α -carbon bead model and a natural protein structure requires certain assumptions because of the difference in the chain (all atom versus bead representation) and the number of amino acids (the protein L model has fewer beads in some of the turn regions). The CE tool was particularly applicable for our purposes for two reasons. First, it compares only the α -carbon positions of the two structures when calculating the structural alignment, and second, the CE algorithm can exclude certain α -carbon positions to align the model and solution structures despite the different lengths of the loop regions. It should be noted that the insertion of gaps in the structural alignment did not result in a spurious alignment. The z-score for the structural alignment was 3.1. This measure indicates that an alignment of that quality with a random structure would occur in 1 in 10^3 times, showing that the protein L bead model has high topological similarity to the protein L natural fold.

Acknowledgments

We acknowledge financial support from UC Berkeley, and a sub-contract award under the National Sciences Foundation Grant No. CHE-0205170. We also thank Nick Fawzi for calculating the RMSD between native states of the protein L model and the experimental structure.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Appendix A: Solution of UIN mechanism

For the two-step reversible mechanism given in equation 1 we have the following differential equations describing the time rate of change in concentration of each species,

$$\frac{d[U]}{dt} = -k_1[U] + k_{-1}[I]$$

$$\frac{d[I]}{dt} = k_1[U] - k_{-1}[I] - k_2[I] + k_{-2}[N] \quad (5)$$

$$\frac{d[N]}{dt} = k_2[I] - k_{-2}[N]$$

These set of coupled first-order differential equations can be straightforwardly solved by a Laplace transform, given by

$$\mathcal{L}\{\dots\} = \int_0^\infty \dots e^{-st} dt \quad (6)$$

Taking the Laplace transform of the differential equations we have

$$\begin{aligned} s\mathcal{L}\{[U]\} - [U]_0 &= -k_1\mathcal{L}\{[U]\} + k_{-1}\mathcal{L}\{[I]\} \\ s\mathcal{L}\{[I]\} - [I]_0 &= -k_1\mathcal{L}\{[U]\} - k_{-1}\mathcal{L}\{[I]\} - k_2\mathcal{L}\{[I]\} + k_{-2}\mathcal{L}\{[N]\} \\ s\mathcal{L}\{[N]\} - [N]_0 &= -k_2\mathcal{L}\{[I]\} - k_{-2}\mathcal{L}\{[N]\} \end{aligned} \quad (7)$$

where $[U]_0$, $[I]_0$, and $[N]_0$ are the initial concentrations at time $t = 0$. Rearranging gives the set of linear equations

$$\begin{aligned} (s + k_1)\mathcal{L}\{[U]\} - k_{-1}\mathcal{L}\{[I]\} &= [U]_0 \\ -k_1\mathcal{L}\{[U]\} + (s + k_2 + k_{-1})\mathcal{L}\{[I]\} - k_{-2}\mathcal{L}\{[N]\} &= [I]_0 \\ -k_2\mathcal{L}\{[I]\} + (s + k_{-2})\mathcal{L}\{[N]\} &= [N]_0 \end{aligned} \quad (8)$$

In matrix form these equations can be expressed as

$$\begin{pmatrix} s + k_1 & -k_{-1} & 0 \\ -k_1 & s + k_{-1} + k_2 & -k_{-2} \\ 0 & -k_2 & s + k_{-2} \end{pmatrix} \begin{pmatrix} \mathcal{L}\{[U]\} \\ \mathcal{L}\{[I]\} \\ \mathcal{L}\{[N]\} \end{pmatrix} = \begin{pmatrix} [U]_0 \\ [I]_0 \\ [N]_0 \end{pmatrix} \quad (9)$$

For the mechanism we are interested in investigating here, we have the initial conditions that $[I]_0 = [N]_0 = 0$. Using Cramer's rule we can express the solutions to the above set of equations (Strang 1988):

$$\mathcal{L}\{[U]\} = \frac{\begin{vmatrix} [U]_0 & -k_{-1} & 0 \\ 0 & s + k_{-1} + k_2 & -k_{-2} \\ 0 & -k_2 & s + k_{-2} \end{vmatrix}}{\begin{vmatrix} s + k_1 & -k_{-1} & 0 \\ -k_1 & s + k_{-1} + k_2 & -k_{-2} \\ 0 & -k_2 & s + k_{-2} \end{vmatrix}} \quad (10a)$$

$$\mathcal{L}\{[I]\} = \frac{\begin{vmatrix} s + k_1 & [U]_0 & 0 \\ -k_{-1} & 0 & -k_{-2} \\ 0 & 0 & s + k_{-2} \end{vmatrix}}{\begin{vmatrix} s + k_1 & -k_{-1} & 0 \\ -k_1 & s + k_{-1} + k_2 & -k_{-2} \\ 0 & -k_2 & s + k_{-2} \end{vmatrix}} \quad (10b)$$

Finding the solutions to the determinants and simplifying gives

$$\mathcal{L}\{[U]\} = [U]_0 \left[\frac{(s^2 + k_{-1}k_{-2})}{s(s - r_1)(s - r_2)} + \frac{(k_2 + k_{-1} + k_{-2})}{(s - r_1)(s - r_2)} \right] \quad (11a)$$

$$\mathcal{L}\{[I]\} = [U]_0 \left[\frac{k_1(s + k_{-2})}{s(s - r_1)(s - r_2)} \right]; \quad (11b)$$

where r_1 and r_2 are given by

$$\begin{aligned} r_1 &= -\frac{1}{2}(k_1 + k_2 + k_{-1} + k_{-2}) + \frac{1}{2}[(k_1 + k_2 + k_{-1} + k_{-2})^2 \\ &\quad - 4(k_1k_2 + k_1k_{-2} + k_{-1}k_{-2})]^{1/2} \end{aligned} \quad (12a)$$

and

$$\begin{aligned} r_2 &= -\frac{1}{2}(k_1 + k_2 + k_{-1} + k_{-2}) - \frac{1}{2}[(k_1 + k_2 + k_{-1} + k_{-2})^2 \\ &\quad - 4(k_1k_2 + k_1k_{-2} + k_{-1}k_{-2})]^{1/2}. \end{aligned} \quad (12b)$$

Taking the inverse Laplace transforms gives us solutions for [U] and [I] as a function of time:

$$[U] = [U]_0 \left(\frac{k_{-1}k_{-2}}{r_1r_2} + \left[\frac{r_1^2 + k_{-1}k_{-2}}{r_1^2 - r_1r_2} + \frac{k_2 + k_{-1} + k_{-2}}{r_1 - r_2} \right] e^{r_1t} + \left[\frac{r_2^2 + k_{-1}k_{-2}}{r_2^2 - r_1r_2} + \frac{k_2 + k_{-1} + k_{-2}}{r_1 - r_2} \right] e^{r_2t} \right) \quad (13a)$$

$$[I] = [U]_0 k_1 \left(\frac{k_{-2}}{r_1r_2} + \frac{r_1 + k_{-2}}{r_1^2 - r_1r_2} e^{r_1t} + \frac{r_2 + k_{-2}}{r_2^2 - r_1r_2} e^{r_2t} \right) \quad (13b)$$

The condition of detailed balance gives the final equation for [N],

$$[N] = [U]_0 - [U] - [I] \quad (14)$$

Thus, we obtain solutions for [U], [I], and [N] as functions of time.

References

- Abkevich, V.I., Gutin, A.M., and Shakhnovich, E.I. 1994. Specific nucleus as the transition state for protein folding—Evidence from the lattice model. *Biochemistry* **33**: 10026–10036.
- Alm, E., Morozov, A.V., Kortemme, T., and Baker, D. 2002. Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* **322**: 463–476.
- Andersen, H. 1983. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **52**: 24–34.
- Brown, S., Fawzi, N.J., and Head-Gordon, T. 2003. Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci.* **100**: 10712–10717.
- Bryngelson, J.D. and Wolynes, P.G. 1989. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **93**: 6902–6915.
- Daggett, V. and Fersht, A. 2003a. The present view of the mechanism of protein folding (review). *Nat. Rev. Mol. Cell Biol.* **4**: 497–502.
- . 2003b. Is there a unifying mechanism for protein folding? (Review). *Trends Biochem. Sci.* **28**: 18–25.
- Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T., and Shakhnovich, E.S. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* **108**: 334–350.
- Ferguson, D.M. and Garrett, D.G. 1999. Simulated annealing—Optimal histogram methods. *Monte Carlo Methods Chem. Phys.* **105**: 311–336.
- Ferrenberg, A.M. and Swendsen, R.H. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63**: 1195–1198.
- Fersht, A.R. 1997. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**: 3–9.
- Gruebele, M. 2002a. An intermediate seeks instant gratification. *Nat. Struct. Biol.* **9**: 154–155.
- . 2002b. Protein folding: The free energy surface. *Curr. Opin. Struct. Biol.* **12**: 161–168.
- Gu, H.D., Yi, Q.A., Bray, S.T., Riddle, D.S., Shiao, A.K., and Baker, D. 1995. A phage display system for studying the sequence determinants of protein folding. *Protein Sci.* **4**: 1108–1117.
- Gu, H.D., Kim, D., and Baker, D. 1997. Contrasting roles for symmetrically disposed β -turns in the folding of a small protein. *J. Mol. Biol.* **274**: 588–596.
- Guo, Z.Y. and Thirumalai, D. 1995. Kinetics of protein folding—Nucleation mechanism, time scales, and pathways. *Biopolymers* **36**: 83–102.
- . 1996. Kinetics and thermodynamics of folding of a de novo designed four-helix bundle protein. *J. Mol. Biol.* **263**: 323–343.
- Guo, Z.Y., Thirumalai, D., Honeycutt, J.D. 1992. Folding kinetics of proteins—A model study. *J. Chem. Phys.* **97**: 525–535.
- Head-Gordon, T. and Brown, S. 2003. Minimalist models for protein folding and design. *Curr. Opin. Struct. Biol.* **13**: 160–167.
- Honeycutt, J.D. and Thirumalai, D. 1990. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci.* **87**: 3526–3529.
- Karanicolas, J. and Brooks, C.L. 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11**: 2351–2361.
- Kim, D.E., Fisher, C., and Baker, D. 2000. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**: 971–984.
- Krantz, B.A., Mayne, L., Rumbley, J., Englander, S.W., and Sosnick, T.R. 2002. Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324**: 359–371.
- Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., and Kollman, P.A. 1995. Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **16**: 1339–1350.
- McCallister, E.L., Alm, E., and Baker, D. 2000. Critical role of β -hairpin formation in protein G folding. *Nat. Struct. Biol.* **7**: 669–673.
- Myers, J.K. and Oas, T.G. 2002. Mechanisms of fast protein folding (review). *Annu. Rev. Biochem.* **71**: 783–815.
- Nauli, S., Kuhlman, B., and Baker, D. 2001. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **8**: 602–605.
- Nymeyer, H., Garcia, A.E., and Onuchic, J.N. 1998. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci.* **95**: 5921–5928.
- Onuchic, J.N., Luthey-Schulten, Z., and Wolynes, P.G. 1997. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**: 545–600.
- Ozkan, S.B., Dill, K.A., and Bahar, I. 2002. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.* **11**: 1958–1970.
- Park, S.H., Oneil, K.T., and Roder, H. 1997. An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry* **36**: 14277–14283.
- Park, S.H., Shastry, M.C.R., and Roder, H. 1999. Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. *Nat. Struct. Biol.* **6**: 943–947.
- Parker, M.J. and Marqusee, S. 1999. The cooperativity of burst phase reactions explored. *J. Mol. Biol.* **293**: 1195–1210.
- Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**: 985–994.
- Plaxco, K.W., Millett, I.S., Segel, D.J., Doniach, S., and Baker, D. 1999. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.* **6**: 554–556.
- Qin, Z., Ervin, J., Larios, E., Gruebele, M., and Kihara, H. 2002. Formation of a compact structured ensemble without fluorescence signature early during ubiquitin folding. *J. Phys. Chem.* **106**: 13040–13046.
- Sali, A., Shakhnovich, E., and Karplus, M. 1994. How does a protein fold. *Nature* **369**: 248–251.
- Sanchez, I.E. and Kiefhaber, T. 2003a. Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol.* **325**: 367–376.
- . 2003b. Hammond behavior versus ground state effects in protein folding: Evidence for narrow free energy barriers and residual structure in unfolded states. *J. Mol. Biol.* **327**: 867–884.
- Scalley, M.L., Yi, Q., Gu, H.D., McCormack, A., Yates, J.R., and Baker, D. 1997. Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry* **36**: 3373–3382.
- Shimada, J. and Shakhnovich, E.I. 2002. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci.* **99**: 11175–11180.
- Shindyalov, I.N. and Bourne, P.E. 1998. Protein structure alignment by incre-

- mental combinatorial extension (Ce) of the optimal path. *Protein Eng.* **11**: 739–747.
- Sorenson, J.M. and Head-Gordon, T. 1999. Redesigning the hydrophobic core of a model β -sheet protein: Destabilizing traps through a threading approach. *Proteins* **37**: 582–591.
- . 2000. Matching simulation and experiment: A new simplified model for simulating protein folding. *J. Comput. Biol.* **7**: 469–481.
- . 2002a. Toward minimalist models of larger proteins: A ubiquitin-like protein. *Proteins* **46**: 368–379.
- . 2002b. Protein engineering study of protein L by simulation. *J. Comput. Biol.* **9**: 35–54.
- Speed, M.A., Morshead, T., Wang, D.I.C., and King, J. 1997. Conformation of P22 tailspike folding and aggregation intermediates probed by monoclonal antibodies. *Protein Sci.* **6**: 99–108.
- Strang, G. 1988. *Linear algebra and its applications*. Harcourt Brace Jovanovich, Inc., New York.
- Wagner, C. and Kiefhaber, T. 1999. Intermediates can accelerate protein folding. *Proc. Natl. Acad. Sci.* **96**: 6716–6721.