THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 285, NO. 24, pp. 18376–18384, June 11, 2010


# Protein β-Sheet Nucleation Is Driven by Local Modular Formation*[S]

Received for publication, March 5, 2010 Published, JBC Papers in Press, April 10, 2010, DOI 10.1074/jbc.M110.120824

**Brent Wathen**[1] **and Zongchao Jia**[2]

*From the Department of Biochemistry, Queen's University, Kingston, Ontario K7L 3N6, Canada*

**Despite its central role in the protein folding process, the specific mechanism(s) behind β-sheet formation has yet to be determined. For example, whether the nucleation of β-sheets, often containing strands separated in sequence by many residues, is local or not remains hotly debated. Here, we investigate the initial nucleation step of β-sheet formation by performing an analysis of the smallest β-sheets in a non-redundant dataset on the grounds that the smallest sheets, having undergone little growth after nucleation, will be enriched for nucleating characteristics. We find that the residue propensities are similar for small and large β-sheets as are their interstrand pairing preferences, suggesting that nucleation is not primarily driven by specific residues or interacting pairs. Instead, an examination of the structural environments of the two-stranded sheets shows that virtually all of them are contained in single, compact structural modules, or when multiple modules are present, one or both of the chain termini are involved. We, therefore, find that β-nucleation is a local phenomenon resulting either from sequential or topological proximity. We propose that β-nucleation is a result of two opposite factors; that is, the relative rigidity of an associated folding module that holds two stretches of coil close together in topology coupled with sufficient chain flexibility that enables the stretches of coil to bring their backbones in close proximity. Our findings lend support to the hydrophobic zipper model of protein folding (Dill, K. A., Fiebig, K. M., and Chan, H. S. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 1942–1946). Implications for protein folding are discussed.**

β-Sheets, a prominent part of protein architecture first predicted almost 60 years ago (1, 2), provide an important avenue for folding proteins to maximize the formation of backbone hydrogen bonds, particularly within their cores, that significantly enhance their structural stability. Unlike other prominent types of protein structure such as helices and turns, β-sheets need not be local in nature, with many involving residues separated in sequence by hundreds of amino acids. This non-local nature dramatically complicates the study and prediction of β-sheet formation, as the search for interacting strands becomes a whole-sequence endeavor. With the formation of non-native β-structures now implicated in such debilitating diseases as Alzheimer and Parkinson (3), unraveling the mechanisms of sheet formation has become more important than ever.

Toward this end, considerable research has been done on β-architecture, particularly in the past two decades. With the explosive growth in determined protein structures, statistical analyses of β-architecture have identified intrinsic β-forming propensities for each amino acid (4–8), including at strand ends (9) where a transition occurs from sheet to non-sheet, and structural studies have investigated such features as β-bulges (10), sheet twisting (11), and strand topology (12). Data mining has also revealed interstrand residue pairing preferences that may assist with structure prediction (6, 7, 13–18), whereas the importance of hydrophobicity for sheet formation has been noted (19). Despite these advances, however, the details of sheet formation remain poorly understood. At an elementary level, sheet formation is thought to involve an initial nucleation step wherein two distinct stretches of a polypeptide chain associate to form a new β-seed and subsequent growth steps either from further associations between the two initial strands or from the association of additional strands from other parts of the polypeptide chain. The hydrophobic zipper model of Dill *et al.* (20), for example, describes sheet growth based on local hydrophobic interactions beginning with an initial hydrophobic interaction that serves to limit subsequent conformational searches. Correct strand-strand registration is undoubtedly critical for avoiding excessive non-native sheet formation.

Experimentally, β-sheet formation, like protein folding in general, is a challenging process to study, as the time frames involved are rapid, and few intermediates are isolatable. Initial studies of small peptides and protein fragments rich in residues with high β-sheet propensities were hampered by aggregation (21, 22). However, the discovery of small, soluble peptides that are able to form two- and three-stranded anti-parallel β-sheets dramatically altered the β-sheet experimental landscape (for reviews, see Refs. 23–25). Since then numerous experimental (26–33), theoretical (34), and computational (35–38) studies have appeared, with most suggesting that β-hairpin formation depends on some combination of turn formation and cross-strand hydrophobic stabilization. One widely held view is that early turn formation controls the kinetics of β-hairpin formation, whereas the subsequent creation of a cross-strand hydrophobic core provides the thermodynamic stabilization (27–30, 33–35, 37). This view is not, however, universal. Other experimental and computational studies report an opposing β-sheet formation mechanism, one where nonspe-

* This work was supported by the Canadian Institutes of Health Research.
[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Tables ST1–ST5 and Fig. SF1.
[1] Recipient of Natural Science and Engineering Research Council of Canada graduate studentship.
[2] Canada Research Chair in Structural Biology. To whom correspondence should be addressed: Dept. of Biochemistry, Queen's University, Kingston, Ontario, Canada K7L 3N6, Canada. Tel.: 613-533-6277; E-mail: jia@queensu.ca.

cific hydrophobic collapse occurs first followed by a rate-limiting search through a series of collapsed states for the correct native alignment (32, 36).

In this work we perform statistical analyses and data mining on a non-redundant set of protein structures to investigate the initial nucleation step in β-sheet formation. To focus our analysis on nucleating features, we have investigated the smallest β-sheets, particularly those with just two strands that have undergone nucleation but little if any subsequent growth. Surprisingly, despite using a large dataset, we find that essentially all two-stranded β-sheets undergo topologically local nucleation; virtually all such sheets are either contained in the same folding module or involve one (or both) termini of a protein chain. Implications for both the folding of larger β-sheets and proteins in general are discussed.

## MATERIALS AND METHODS

A non-redundant dataset of protein structures with 40 or more residues was assembled using the PDB-REPRDB server (39) consisting of protein chains from the RCSB Protein Data Bank, release #2007_01_21. Only structures solved by x-ray crystallography that have resolutions less than or equal to 2.0 Å and R-factors less than or equal to 0.25 were included. Those with chain breaks or with missing non-hydrogen atoms were excluded as were membrane proteins and polypeptide chains that are part of larger structural complexes. To ensure non-redundancy, only one representative was chosen for the dataset from among those protein chains that had sequence identities greater than 30% or structural alignments less than 10 Å. In addition, several chains were found to have the same fold by visual inspection; in these cases, only one representative was retained in the dataset. In all, the dataset consisted of 1334 protein chains, listed in supplemental Table ST1.

The DSSP program (40) was used to identify the secondary structure of each residue in the dataset. Two-stranded sheets with six or more residues and three-or-more-stranded sheets with seven or more residues were analyzed in this study (supplemental Table ST2). A number of two-stranded β-sheets were found to be closely associated with larger β-sheets; to avoid the possibility that the nucleation of these smaller sheets was influenced by these associations, these smaller sheets were also excluded from analysis. In all, 439 2-stranded sheets, and 2114 three-or-more-stranded sheets were included in our analysis.

β-Propensity values $P_r$ for each amino acid $r$ were calculated using the Chou-Fasman propensity formula (4),

$$P_r = [S_r/S_{all}]/[D_r/D_{all}] \quad \text{(Eq. 1)}$$

where $S_r$ and $S_{all}$ are the number of $r$ residues and of all residues in strands in the dataset, and $D_r$ and $D_{all}$ are the total number of $r$ residues and of all residues in the dataset. Separate propensity values were calculated analogously for two-stranded β-sheets, two-stranded β-sheets with eight or less residues, two-stranded β-sheets with nine or more residues, three-or-more-stranded β-sheets, and for the edge strands in three-or-more-stranded β-sheets.

Interstrand pairing preferences between residues $a$ and $b$ in two-stranded β-sheets, $P_{a,b}$, were calculated using

**TABLE 1**

**Residue propensities for various β-sheet environments**

Propensities values are calculated as described under "Materials and Methods."

| Residue | All sheets | Two-stranded sheets | | | Three-or-more-stranded sheets | |
|---|---|---|---|---|---|---|
| | | All | Smallest[a] | Largest[b] | All | Edge strands |
| Ala | 0.77 | 0.61 | 0.62 | 0.58 | 0.78 | 0.68 |
| Val | 1.94 | 1.73 | 1.66 | 1.84 | 1.95 | 1.75 |
| Leu | 1.18 | 0.94 | 0.97 | 0.90 | 1.19 | 1.02 |
| Ile | 1.77 | 1.47 | 1.57 | 1.30 | 1.79 | 1.62 |
| Phe | 1.42 | 1.38 | 1.37 | 1.40 | 1.42 | 1.37 |
| Tyr | 1.43 | 1.44 | 1.44 | 1.44 | 1.43 | 1.28 |
| Met | 1 | 1.06 | 1.05 | 1.07 | 1.00 | 0.79 |
| Cys | 1.24 | 1.16 | 1.09 | 1.27 | 1.24 | 1.12 |
| Trp | 1.31 | 1.44 | 1.06 | 2.07 | 1.31 | 1.31 |
| His | 0.91 | 0.91 | 0.84 | 1.02 | 0.91 | 0.93 |
| Glu | 0.72 | 0.87 | 0.70 | 1.15 | 0.71 | 0.98 |
| Gln | 0.77 | 0.96 | 0.93 | 1.01 | 0.76 | 0.91 |
| Lys | 0.79 | 1.05 | 1.01 | 1.12 | 0.77 | 1 |
| Arg | 0.93 | 1.08 | 1.13 | 1.00 | 0.92 | 1.04 |
| Ser | 0.82 | 0.78 | 0.84 | 0.67 | 0.82 | 0.94 |
| Thr | 1.18 | 1.49 | 1.53 | 1.42 | 1.17 | 1.25 |
| Asp | 0.54 | 0.64 | 0.65 | 0.61 | 0.54 | 0.58 |
| Asn | 0.6 | 0.67 | 0.74 | 0.56 | 0.60 | 0.6 |
| Gly | 0.63 | 0.49 | 0.49 | 0.48 | 0.63 | 0.55 |
| Pro | 0.44 | 0.74 | 0.86 | 0.55 | 0.43 | 0.62 |
| Correlations[c] | | | | | 0.89 | 0.94 |

[a] Calculated for two-stranded sheets with eight or fewer residues.
[b] Calculated for two-stranded sheets with nine or more residues.
[c] Correlations are with the residue propensities for all two-stranded β-sheets.

$$P_{a,b} = I_{a,b}/E_{a,b} \quad \text{(Eq. 2)}$$

where $I_{a,b}$ and $E_{a,b}$ are the actual and expected number of times residues $a$ and $b$ are aligned opposite one another on adjacent two-stranded β-strands. $E_{a,b}$ is given by,

$$E_{a,b} = [N_a{*}N_b/N^2]{*}I_{all} \quad \text{(Eq. 3)}$$

where $N_a$ and $N_b$ are the number of times residues $a$ and $b$ are found in two-stranded β-sheets, $N$ is the total number of residues in two-stranded β-sheets, and $I_{all}$ is the total number of interstrand pairings between strands in two-stranded sheets. Visual inspection of the sheets in this study was performed using Pymol (DeLano Scientific, Palo Alto, CA).

## RESULTS

*Residue Propensities within Two-stranded β-Sheets*—Residue propensities for various β-environments were calculated as described under "Materials and Methods." The non-redundant dataset employed (hereafter referred to as "the dataset") contains 1334 protein chains containing a total of 2553 β-sheets (see the supplemental Tables ST1 and ST2 for lists of protein chains and β-sheets used in this study). Residue β-sheet propensities are highly correlated with those reported in other studies (Tables 1, supplemental ST3), suggesting that our data base is not unduly biased. Edge and interior strand propensity values have also been calculated separately (Table 1).

Given a two-step model for β-sheet formation consisting of initial nucleation followed by subsequent growth, we have calculated propensity values for two-stranded and three-or-more-stranded β-sheets separately (Table 1) on the assumption that the former by necessity will contain their nucleation seeds and, hence, be enriched for nucleating residues. In addition, given that two-stranded β-sheets consist solely of edge strands, we have separately calculated edge and interior pro-

**TABLE 2**

**Interstrand residue pairing preferences from two-stranded β-sheets in the data set**

Values are the *z* scores measuring the differences between actual and expected frequency counts as described under "Materials and Methods." Values larger than 2.5 are shown in bold and double underlined.

| | Ala | Val | Leu | Ile | Phe | Tyr | Met | Cys | Trp | His | Glu | Gln | Lys | Arg | Ser | Thr | Asp | Asn | Gly | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 0.81 | | | | | | | | | | | | | | | | | | | |
| Val | 1.14 | 1.49 | | | | | | | | | | | | | | | | | | |
| Leu | 1.38 | 1.31 | 1.37 | | | | | | | | | | | | | | | | | |
| Ile | 0.73 | 1.22 | 1.18 | 1.86 | | | | | | | | | | | | | | | | |
| Phe | 0.59 | 1.02 | 1.02 | 1.28 | 1.33 | | | | | | | | | | | | | | | |
| Tyr | 1.37 | 1.12 | 1.00 | 0.93 | 1.18 | 0.88 | | | | | | | | | | | | | | |
| Met | 2.01 | 1.09 | 0.73 | 0.81 | 0.65 | 1.52 | 0.99 | | | | | | | | | | | | | |
| Cys | 1.11 | 0.31 | 1.44 | 1.45 | 0.71 | 0.77 | 0.82 | **13.4** | | | | | | | | | | | | |
| Trp | 1.78 | 0.75 | 1.16 | 0.75 | 1.23 | 1.86 | 1.12 | **3.24** | 1.28 | | | | | | | | | | | |
| His | 1.93 | 0.82 | 1.57 | 0.79 | 1.05 | 1.13 | 0.00 | 0.00 | 0.00 | 2.30 | | | | | | | | | | |
| Glu | 0.19 | 0.77 | 1.11 | 0.90 | 0.60 | 0.84 | 1.05 | 0.52 | 1.19 | 0.45 | 0.53 | | | | | | | | | |
| Gln | 0.63 | 0.84 | 0.92 | 0.83 | 0.92 | 1.32 | 0.70 | 0.00 | 0.79 | 1.13 | 1.33 | 0.99 | | | | | | | | |
| Lys | 0.68 | 0.87 | 0.68 | 0.93 | 0.66 | 1.19 | 0.81 | 0.83 | 1.49 | 0.87 | **2.57** | 1.14 | 0.49 | | | | | | | |
| Arg | 0.86 | 0.69 | 0.73 | 1.16 | 1.48 | 1.59 | 1.35 | 0.00 | 0.89 | 0.73 | 2.34 | 0.40 | 0.60 | 0.82 | | | | | | |
| Ser | 0.77 | 0.68 | 0.39 | 0.70 | 1.32 | 0.56 | 0.79 | 2.39 | 1.05 | 0.85 | 0.89 | 1.86 | 1.45 | 1.32 | 1.40 | | | | | |
| Thr | 1.09 | 0.97 | 0.35 | 0.76 | 0.93 | 0.77 | 0.71 | 0.23 | 0.32 | 1.38 | 0.75 | 1.00 | 0.78 | 0.64 | 1.21 | **2.65** | | | | |
| Asp | 0.86 | 0.61 | 0.42 | 0.19 | 0.56 | 0.15 | 0.95 | 0.00 | 0.72 | **2.74** | 1.01 | 1.35 | **2.78** | 1.73 | 1.27 | 0.91 | 0.61 | | | |
| Asn | 0.75 | 0.72 | 0.86 | 0.49 | 1.00 | 0.20 | 0.42 | 0.69 | 0.47 | 0.90 | 1.06 | 2.20 | 1.36 | 1.04 | 0.88 | 1.37 | 1.20 | 1.75 | | |
| Gly | 1.44 | 1.11 | 0.94 | 0.90 | 0.86 | 1.00 | 1.21 | 1.50 | 1.03 | 0.33 | 0.71 | 0.64 | 0.25 | 0.28 | 0.81 | 1.31 | 2.14 | 1.02 | 0.74 | |
| Pro | 1.04 | 0.76 | 1.36 | 0.59 | 1.59 | 1.25 | 1.65 | 1.09 | 1.12 | 0.00 | 0.84 | 0.47 | 0.27 | 1.72 | 1.32 | 0.99 | 0.64 | 1.66 | 1.21 | 0.88 |

pensity values for the three-or-more-stranded sheets for comparison. In all, there are 439 2-stranded β-sheets in the dataset. High correlations between residue propensities for 2-stranded and 3-or-more-stranded β-sheets ($r = 0.89$) and between residues propensities for 2-stranded and edge strands ($r = 0.94$) are observed. That the residue propensities for two-stranded β-sheets are more closely correlated to those for edge strands in larger sheets is not surprising given that both strands in a two-stranded sheet must not only code for nucleation but, being edge strands, also discourage further strand associations. Nonetheless, several residues have moderately higher propensities for 2-stranded sheets than for edge strands in larger sheets, including (with percent increases shown in parentheses) Met (34%), Thr (19%), Tyr (13%), Asn (12%), and, unexpectedly, Pro (19%) (Table 1). A few have lower preferences, particularly Ser (-17%). Overall, however, no distinguishing trends are evident in the two-stranded residue propensities that might account for β-sheet nucleation.

To further focus our search for nucleating elements within β-sheets, we calculated residue propensity values for two-stranded sheets with eight or fewer residues separately from those with more than eight residues, again on the assumption that the smaller ones consist almost exclusively of their nucleation seeds (Table 1). In total there are 328 2-stranded β-sheets with 8 or fewer residues in the dataset and 111 2-stranded sheets with more than 8 residues. Notably, few residues show higher propensity for being in the smaller 2-stranded sheets than in the larger ones; only Ile (1.57 *versus* 1.30), Ser (0.84 *versus* 0.67), Asn (0.74 *versus* 0.56), and surprisingly, Pro (0.86 *versus* 0.55) have elevated propensities for the smallest sheets. Moreover, the elevated propensities for Ile and Ser are very similar to their propensities for edge strands in larger β-sheets, suggesting that the variations observed may simply be random fluctuations. Again, we do not find any clear trends that might be related to β-sheet nucleation.

*Interstrand Pairing Preferences within Two-stranded β-Sheets*—As β-nucleation involves an interaction between two separate parts of a polypeptide chain, we have also checked for specific cross-strand pairing preferences between spatially neighboring residues in two-stranded β-sheets. Previous studies of pairing preferences have confirmed that cross-strand residue pairings are non-random, and specific pairing preferences for various structural environments have been identified and rationalized (6, 7, 13–18). In the present study we have restricted our analysis to cross-strand pairing preferences found on two-stranded β-sheets to focus on β-nucleating characteristics. Preferences have been normalized against expected values for the specific residue distributions in two-stranded sheets, as described under "Materials and Methods." As seen in Table 2, general preferences exist between pairs of hydrophobic residues, and between pairs of hydrophilic residues. Particularly high preferences are seen between pairs of Cys (13.40) and Thr (2.65) residues as well as between Asp-His pairs (2.74), Asp-Lys pairs (2.78), Glu-Lys pairs (2.57), and oppositely charged residues in general, whereas notably low preferences are seen for pairings of hydrophobic and hydrophilic residues. Cys-Trp pairings also occur more often than expected (3.24), but as there are only 7 instances of this pairing in the dataset, they cannot play a substantial role in β-nucleation. Direct comparison of these pairing preferences with those from larger β-architectures is hampered by the fact that pairings from larger sheets necessarily involve at least one interior strand, and interior strands are known to have different residue propensities in general (6). Nevertheless, all of the favored cross-strand pairings for two-stranded β-sheets have previously been reported to have high occurrence frequencies across β-sheets in general (6, 15, 18), suggesting little enrichment exists that would account for β-nucleation specifically.

We also separately calculated the cross-strand pairing preferences for two-stranded β-sheets with eight or less, or more

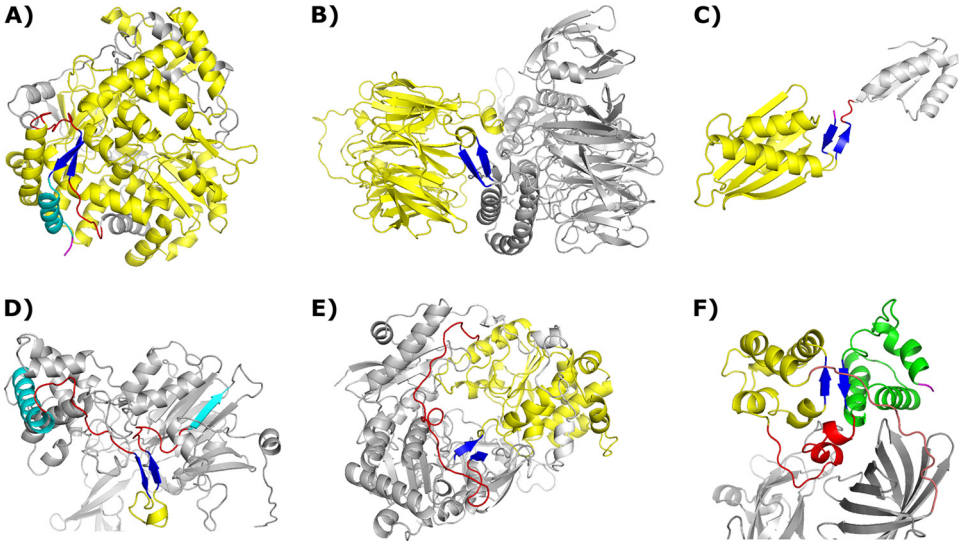


FIGURE 1. **Representative two-stranded β-sheets in the dataset.** Common coloration: two-stranded β-sheets are shown in *blue*, intervening residues are shown in *yellow*, relevant stretches of coil are shown in *red*, and relevant chain termini are shown in *magenta*. *A*, shown is a two-stranded β-sheet encompassing multiple modules. This sheet has the largest strand separation (499 residues), and the intervening residues compose the majority of the folded protein. One strand of the sheet is in the C-terminal region followed by a final helix (*cyan*); the other occurs in the midst of a long coil region (PDB code 1HSS). *B*, module capping, with intervening residues forming a large β-propeller, is shown (PDB code 1K32). *C*, module capping involving chain termini is shown (PDB code 2CVE). *D*, shown is a two-stranded β-sheet in a stretch of extended coil transiting between two distinct modules. The final helix of one module and the initial strand of the next are shown in *cyan* (PDB code 1PNK). *E*, shown is a two-stranded β-sheet with an intervening structural module that has different entry/exit points. In this case a long extended coil region connects the entry and exit points together, allowing for sheet formation (PDB code 1DMR). *F*, shown is the only case of a two-stranded β-sheet composed of strands in different structural modules, not involving a chain terminus. The second intervening module is shown in *green*. The coil region joining the two intervening modules is shown in *red*, and the final exiting coil stretch joining these modules to the rest of the protein is shown in *salmon* (PDB code 1PBY).

than eight, residues (supplemental Table ST4), again under the assumption that the smallest sheets should be particularly enriched for nucleating characteristics. In this case, however, the small number of larger two-stranded sheets in the dataset leads to a high signal-to-noise ratio in the pairing preferences for these sheets, resulting in considerable variation between these two groups. Nevertheless, we identify those pairings that satisfy the following criteria:

$$P_{small} > 1.5 \quad \text{(Eq. 4)}$$

$$P_{small} - P_{large} > 1.0 \quad \text{(Eq. 5)}$$

$$F_{small} > 12 \quad \text{(Eq. 6)}$$

where $P_{small}$ and $P_{large}$ are the propensities of a pairing to be in the smallest and largest two-stranded β-sheets, and $F_{small}$ is the frequency with which a pairing is found in the smallest two-stranded sheets. This last condition removes rare pairings from consideration, as their scarcity prevents them from playing a significant role in nucleation. Cross-strand pairings satisfying these criteria included (with $P_{small}$ and $P_{large}$ values in parentheses): Asp-His (3.99/0.86), Asp-Lys (3.62/1.39), Asp-Gly (2.69/1.12), Pro-Arg (2.06/0.60), Leu-Leu (1.91/0.26), Gly-Ala (1.90/0.59), Tyr-Ala (1.80/0.60), Asn-Lys (1.76/0.53), and Phe-Ile (1.72/0.39).

*Strands in Two-stranded β-Sheets*—We examined both the sequential relationship between the strands in two-stranded β-sheets and the structural environment in which these sheets are found for distinguishing features that might elucidate β-nucleation. The majority of sheets in the dataset (273 of439) have strands that are separated by 10 residues or less, including 207, which is separated by at most 5 residues (see supplemental Table ST5 for a complete characterization of all two-stranded β-sheets in the dataset). All but two of the 2-stranded β-sheets with 10 or fewer intervening residues are part of β-hairpins, with either tight turns or somewhat more extended turns between the two strands, as expected given the short length of intervening coil. Nucleation in all of these cases is almost certainly a result of turn formation coupled with chain diffusion that produces favorable interstrand interactions akin to what is proposed in models such as the hydrophobic zipper model of Dill *et al.* (20). The two exceptions are both parallel 2-stranded β-sheets with very short intervening stretches of random coil (PDB code 2CAP, chain A, residues 276–287; PDB code 1QBA, residues 523–536); nucleation is undoubtedly a local phenomenon for these sheets also, perhaps involving the formation of an extra turn that aligns the strands in a parallel rather than antiparallel orientation. A further 73 2-stranded β-sheets in the dataset have between 11 and 30 residues separating their stands, and although it would be tenuous at best to describe such strand separations as local, the vast majority of these sheets (69 of 73), including the residues located between their strands, are contained in single, generally compact structural units, or *modules* following Ref. 41. The remaining 4 (PDB code 1PNK, chain B, residues 452–473, Fig. 1; PDB code 2BLF, chain A, residues 237–257; PDB code 1EPT, chain B, residues 24–47; PDB cod 1EPT, chain B, residues 4–23) are also local in nature, having either hairpins or somewhat more extended turns between their strands but occur in the midst of larger stretches of coil that are not properly described as structural modules *per se* because of their extended nature. Most often, the intervening residues between antiparallel strands separated by 11–30 residues adopt antiparallel architectures themselves, which may include small helices, additional β-sheets, or just coil stretches running in tandem. In the case of parallel strands separated by 11–30 residues, the dominant architectures observed are strand-helix-strand and strand-coil-strand motifs.

In total, then, 346 of the 439 2-stranded β-sheets in the dataset (79%) have strands that are separated by 30 residues or less, and all of these are located in single folding modules or, in 4 cases, as compact motifs in stretches of extended coil. Of the remaining 93 2-stranded β-sheets with larger strand separations, 58 are composed of strands that are separated by 31–100

**TABLE 3**

**Structural environments of two-stranded β-sheets with the highest strand separations**

aa, amino acid.

| PDB code[a] | Strands (length (aa))[b] | Separation[c] | Environment | | Single module?[e] | Intervening motif[f] |
|---|---|---|---|---|---|---|
| | | | Strand 1[d] | Strand 2[d] | | |
| 1HHS(A) | 143(4) 646(4) | 499 | X | CT | No | Multiple |
| 2BMO(A) | 13(3) 380(3) | 364 | NT | M | No | Multiple |
| 1K32(A) | 291(4) 652(4) | 357 | Cap | Cap | Yes | β-Propeller |
| 2BJK(A) | 165(9) 507(9) | 333 | M | CT | No | Multiple |
| 1WTJ(B) | 10(4) 327(4) | 313 | NT | CT | No | Multiple |
| 1Y7B(A) | 3(3) 313(3) | 307 | NT | Cap | Yes | β-Propeller |
| 1DMR | 183(3) 438(3) | 252 | Cap | X | Yes* | α/β |
| 1EDT | 28(3) 267(3) | 236 | M | CT | Yes | β-Barrel |
| 2BUM(B) | 5(5) 231(4) | 221 | NT | CT | Yes | β-Ensemble |
| 2GQT(A) | 51(3) 265(3) | 211 | M | CT | No | Multiple |
| 1C1K(A) | 2(3) 197(3) | 192 | NT | CT | No | Multiple |
| 1X7Y(B) | 18(4) 199(4) | 177 | NT | Cap | Yes | α/β |
| 1V5E(A) | 3(3) 169(3) | 163 | NT | Cap | Yes | α/β |
| 2B3Y(A) | 74(4) 234(4) | 156 | Cap | Cap | Yes | α/β |
| 2FA1(B) | 9(4) 154(4) | 141 | NT | CT | Yes | β-Ensemble |
| 1G8L(A) | 27(5) 171(5) | 139 | NT | Cap | Yes | ↑ ↓-Coil |
| 1GNL(A) | 223(4) 364(4) | 137 | Cap | Cap | Yes | α/β |
| 1GV4(A) | 179(3) 317(3) | 135 | Cap | Cap | Yes | α/β |
| 1ESG(B) | 46(3) 183(3) | 134 | M | M | Yes | α/β |
| 1Q1R(A) | 112(3) 248(3) | 133 | Cap | Cap | Yes | α/β |
| 1FN9(A) | 221(3) 356(3) | 132 | M | CT | No | Multiple |
| 1T1E | 14(3) 147(3) | 130 | NT | M | Yes | ↑ ↓ |
| 3GRS | 158(3) 291(3) | 130 | Cap | Cap | Yes | α/β |
| 1JR2(B) | 57(5) 189(5) | 127 | Cap | Cap | Yes | α/β |
| 2GAG(A) | 251(3) 381(3) | 127 | Cap | Cap | Yes | α/β |
| 2C65(B) | 79(3) 208(3) | 126 | Cap | Cap | Yes | α-Ensemble |
| 1HYU(A) | 324(3) 451(3) | 124 | Cap | Cap | Yes | α/β |
| 1YQZ(A) | 115(3) 239(3) | 121 | Cap | Cap | Yes | α/β |
| 1WCW(A) | 39(5) 163(5) | 119 | Cap | Cap | Yes | α/β |
| 1ZK7(A) | 148(3) 269(3) | 118 | Cap | Cap | Yes | α/β |
| 2CVE(A) | 3(3) 121(3) | 115 | NT | Cap | Yes | α/β |
| 1W4X(A) | 222(3) 339(3) | 114 | X | M | Yes | α-Ensemble |
| 1O7E(B) | 51(3) 166(3) | 112 | Cap | M | Yes | α-Ensemble |
| 1FZW(C) | 111(4) 222(4) | 107 | M | M | Yes | α/β |
| 2DQ6(A) | 308(3) 414(3) | 103 | M | M | Yes | α-Ensemble |

[a] The PDB code is followed by the chain ID (in parentheses).
[b] The first residue for each strand is listed followed by the length of each strand (in parentheses).
[c] Number of intervening residues between the strands in the two-stranded β-sheet.
[d] Environments: NT, chain N-terminal region; CT, chain C-terminal region; M, module interior; Cap, module cap, just beyond a structural module; X, extended coil region.
[e] Indicates whether the strands and the intervening residues are part of s single, compact structural module. Yes*, Yes, with additional extended topology to join the module entry/exit points.
[f] Motifs adopted by interstrand residues. Introduced nomenclature: ↑ ↓, antiparallel architecture, culminating in a β-hairpin turn; ↑ ↓, coil: mostly coil antiparallel architecture.

residues. To the best of our judgment, a visual inspection of these sheets reveals that virtually all of them are contained in single folding modules, either in the midst of these modules or, frequently, in linker regions just beyond these modules, a relationship we term module capping (Fig. 1). The only exception to this rule in our estimation is in protein 1PBY, chain A (Fig. 1); here, the 2-stranded β-sheet (residues 29–31 and 112–114, with 80 residues separating the two strands; Fig. 1) appears to be composed of strands from two distinct, although small, folding modules.

Of particular interest with regard to β-sheet formation are two-stranded β-sheets with large separations between their strands, as diffusion is unlikely to play a significant role in their nucleation. In the dataset, only 35 2-stranded sheets have more than 100 residues between their strands (8% of the total); these are listed in Table 3. Strikingly, 13 of the top 16 β-sheets with the largest strand separations involve either the chain N terminus (NT) or the chain C terminus (CT), or both (Fig. 1). In addition, 20 of the 35 involve strands that cap a single structural module, occurring in flexible regions (including linker regions and chain termini) just beyond the module, whereas an additions 6 are embedded in the midst of single modules. Seven clearly span multiple modules, but, notably, in all of these instances the two-stranded sheets in question involve chain termini. Two further sheets defy easy module classification (PDB code 1DMR, residues 183–185 and 438–440; PDB code 1W4X, chain A, residues 222–224 and 339–341). Both cases effectively cap a single module, but unlike the many other instances of module capping, the chain entry and exit points for these modules are not spatially close. To compensate, there is an extended stretch of coil in one case (PDB code 1DMR, residues 408–437, Fig. 1) and a long coil/helix stretch in the other (PDB code 1W4X, residues 225–258, supplemental Fig. SF1) that bring these module entry and exit points together, enabling the creation of the two-stranded sheets in question. In both cases, extensive contacts between these extended stretches and other parts of the protein effectively guide the two strands together.

To summarize, of the 439 2-stranded β-sheets in the dataset, only 35 have strands separated by more than 100 residues. Almost all of the sheets with the largest strand separations involve strands at the chain termini. By inspection, 429 of these sheets are contained in single structural modules, whereas two others (PDB code 1DMR and 1W4X, supplemental Fig. SF1, described above) that are also associated with single structural

**TABLE 4**

**Structural environments of three-or-more-stranded β-sheets with the highest minimal strand separatio**

aa, amino acid.

| PDB code[a] | No. of strands | Min. Sep.[b] | Adjacent strands[c] length (aa) | Environment[d] | | Single module[e] | Intervening motif[f] |
|---|---|---|---|---|---|---|---|
| | | | | Strand 1 | Strand 2 | | |
| 1R9D(A) | 3 | 95 | 449(4), 548(4) | Cap | M | Yes | α-Ensemble |
| 2B0T(A) | 3 | 74 | 247(4), 339(4) | M | Cap | Yes | α-Ensemble |
| 2B3F(A) | 3 | 64 | **119(2), 272(5)** | Cap | Cap | Yes | α/β |
| 1FN9(A) | 3 | 48 | **127(5), 361(4)** | M | CT | No | Multiple |
| 1YNH(B) | 3 | 48 | 14(8), 70(4) | NT | M | Yes | βαβ |
| 2BEM(A) | 3 | 46 | 75(9), 130(8) | M | M | Yes | ↑ ↓-β |
| 1S9R(A) | 3 | 45 | 20(6), 71(4) | NT | M | Yes | βαβ |
| 1XGS(A) | 3 | 42 | **189(3), 276(3)** | M | CT | Yes | ↑ ↓ |
| 2CI3(A) | 3 | 41 | 14(5), 60(5) | NT | M | Yes | βαβ |
| 2HBV(B) | 3 | 37 | 74(4), 115(3) | M | M | Yes | βαβ |
| 1UG6(A) | 8 | 34 | 73(5), 112(7) | M | M | Yes | βαβ |
| 1PAM(A) | 3 | 33 | **586(9), 636(8)** | M | M | Yes | ↑ ↓-β |
| 2GB0(B) | 3 | 32 | **84(2), 145(3)** | Cap | Cap | Yes | α/β |

[a] The PDB code is followed by the chain ID (in parentheses). As defined in Table 3.
[b] Minimum separation (Min. Sep.) refers to the least number of intervening residues between sequentially adjacent pairs of strands in a β-sheet.
[c] Topologically adjacent strands that are candidates for β-nucleation. In most cases these are the minimally separated strands. In a few cases (bold), different pairs of strands are considered better candidates for nucleation, as described in "Strands in Larger β-Sheets."
[d] Environments: NT, chain N-terminal region; CT, chain C-terminal region; M, module interior; Cap, module cap, just beyond a structural module; X, extended coil region.
[e] Indicates whether the strands and the intervening residues are part of s single, compact structural module. Yes*, Yes, with additional extended topology to join the module entry/exit points.
[f] Motifs adopted by interstrand residues. Introduced nomenclature: ↑ ↓, antiparallel architecture, culminating in a β-hairpin turn; ↑ ↓, coil: mostly coil antiparallel architecture.

modules involve extended coil/helical stretches to join the module entry and exit points together spatially. In seven sheets, the intervening residues between their strands clearly span multiple structural modules, but in all of these cases, one or both of the strands involve the chain termini. Finally, just 1 of the 439 2-stranded β-sheets in the dataset (PDB code 1PBY, Fig. 1, described above) appears to be composed of strands contained in two separate structural modules.

*Strands in Larger β-Sheets*—Under the assumption that larger β-sheets form from smaller ones, the overwhelmingly local nature, either in sequence or in space, of the two-stranded β-sheets in the dataset prompted us to examine larger sheets to see if this phenomenon might be universal. More specifically, although larger β-sheets are often composed of strands that are separated by long stretches of intervening residues, we wondered whether all β-sheets with three or more strands contain at least one pair of strands that are contained within the same structural module. Such pairings would be the most likely candidates to initiate β-nucleation because of their sequential and spatial proximity. Notably, of the 2114 β-sheets with three-or-more strands in the dataset (supplemental Table ST2), all but 13 contain at least one pair of strands that are within 30 residues of each other, and none is composed of strands that are all separated by 100 residues or more. We examined the structural features of the 13 β-sheets with the largest minimal strand separations in the dataset (Table 4). All but one of these sheets contains three strands, with the exception being a part of an eight-stranded β-barrel. In 8 of the 13, the strands with minimal separations are found adjacent to one another in their β-sheets, and in all of these cases the strands and their intervening residues are part of the same structural modules, making these pairings strong nucleating candidates for their respective sheets. The minimally separated strands in two further β-sheets are also adjacent in topology and contained in a single structural module, but upon inspection, different pairs of strands in these sheets (PDB code 2B3F, chain A, residues 119–120 and 272–276; PDB code 1XGS, chain A, residues 189–191 and 276–278), also contained in single structural modules, may be better candidates for nucleation because of the more compact nature of the intervening residues between them (supplemental Fig. SF1). The minimally separated strands in the remaining three β-sheets in this set (PDB code 1FN9, chain A; PDB code 1PAM, chain A; PDB code 2GB0, chain B) are not adjacent to one another in their β-sheets. In two of these cases, however, the other possible strand pairings in these sheets are both topologically adjacent and contained in the same structural module (PDB code 1PAM, residues 586–594 and 636–643, supplemental Fig. SF1; PDB code 2GB0, residues 84–85 and 145–147, supplemental Fig. SF1), making them prime candidates to initiate β-formation. In the third case, the other possible strand pairing (PDB code 1FN9, residues 127–131 and 361–364) is separated by a collection of helices and coil regions that is difficult to classify into structural modules. Here, however, as is the case with two-stranded sheets that involve multiple modules, one of the two strands in question is found at the chain termini (supplemental Fig. SF1).

In summary, of the 2114 β-sheets in the dataset with three or more strands, 2101 contain at least one pair of strands separated by at most 30 residues. Without direct examination, it is reasonable to assume that the vast majority of these pairings will be in the same structural modules because of sequence proximity. Moreover, 12 of the 13 β-sheets with minimal strand separations greater than 30 residues contain at least one pair of strands that is topologically adjacent and contained in the same structural module. Finally, as was seen with two-stranded β-sheets, the only example of a larger β-sheet that may lack two strands in the same structural module has one of its strands at the chain termini.

## DISCUSSION

β-Sheet formation remains an enigmatic process despite its importance both to the study of protein folding in general and to the study of several highly debilitating diseases such as Alzheimer and Parkinson (3). Although the basic model of ini-

tial nucleation followed by subsequent growth is conceptually simple, the mechanism by which different parts of a polypeptide chain, separated by dozens or perhaps hundreds of residues, coalesce to nucleate a β-sheet is unclear, particularly when one considers the complexity and diversity in β-sheet architecture seen in many of the protein structures determined to date. Our approach to studying this problem is to simplify; we have focused on only the smallest β-sheets in the dataset under the assumption that these sheets, having undergone nucleation but little if any subsequent growth, will be enriched for nucleating characteristics. Our analysis is divided into two parts, (i) an examination of the residue propensities in two-stranded β-sheets and (ii) an examination of the structural environments of these two-stranded β-sheets, including the structures adopted by the residues between the two strands.

Residue propensity values have long been used to study elements of protein secondary structure (for a recent review, see Ref. 42), including β-architecture (4–9). Propensity values have been reported for a variety of different β-environments, including edge strand propensities (6), interior strand propensities (6), β-breaker propensities at the ends of strands (9), and β-bulge propensities at positions within strands (6), but to our knowledge distinct β-nucleating and β-growth propensities have not been reported. This no doubt stems from the fact that identifying the initial nucleating location(s) of a fully formed β-sheet is not generally possible given that protein folding largely remains a black-box phenomenon. Our examination of the residue propensities for the smallest sheets finds that they differ very little from those for larger sheets, particularly those for edge strands from larger sheets (Table 1). Individual differences do not present themselves as trends among classes of residues. Moreover, several of the elevated propensity values for two-stranded β-sheets are in fact counter-intuitive; both Asn and Pro, two residues not known for participating in β-structures, have elevated occurrence frequencies in the smallest sheets. We attribute this to an increased concentration of β-breakers in the smallest sheets (both Asn and Pro are known β-breakers (9)) and not as a reflection of nucleation capabilities. Overall, we do not find any evidence of unusual residue propensities in the smallest β-sheets that might explain β-nucleation. We also looked to see if there might be specific interstrand pairings that might be responsible for β-nucleation, again by comparing pairing frequencies from the smallest β-sheets with those from larger sheets. Despite different structural environments (interstrand pairings in larger sheets will necessarily include at least one residue on an interior strand), we find that the pairings with elevated frequencies in two-stranded β-sheets are the same ones already reported in the literature as having elevated frequencies across β-sheets in general (6, 15, 18). Moreover, when comparing interstrand pairing frequencies for the smallest two-stranded β-sheets with those for larger two-stranded sheets, five of the top six pairings with the highest discrepancies involve three of the top β-breakers (Asp, Gly, and Pro) (9), again suggesting that differences between smaller and larger two-stranded β-sheets reflect differences in the concentration of β-breakers and not nucleating features. Thus, it appears that the smallest sheets are essentially composed of the same residues in similar arrangements as are larger sheets. We, therefore, conclude that β-nucleation is not driven by particular residues or by specific pairings of residues across strands.

Tertiary interactions have long been thought to play a major role in protein folding. In light of this, we have also examined the structural environments of the two-stranded β-sheets in the dataset. The most striking observation from this examination is the degree to which strands in 2-stranded β-sheets are local in nature, either local in sequence (63% of 2-stranded sheets have 10 or fewer residues between their strands) or, more notably, local in topology. Only 8% of the 2-stranded β-sheets in the dataset have strands separated by more than 100 residues, and of those with the largest strand separations, most involve chain termini. The most telling observation is that almost all two-stranded sheets are associated with a single structural module, either contained within a compact, sequentially contiguous unit or located just beyond such a module (module "capping"). Although one might expect that the majority of two-stranded sheets would indeed be contained in a single structural module, given that most of these sheets have strands that are local in sequence, it is the degree to which this relation holds that is unexpected, with almost no exceptions to this rule. When larger β-sheets with three or more strands are considered, this rule becomes nearly universal; 99.4% of the 2114 larger sheets in the dataset have a pair of strands separated by 30 residues or less, and of the exceptions only one does not have two neighboring strands that are part of a single structural module.

Two-stranded β-sheet nucleation, then, almost always occurs in tandem with the folding of a single structural module. A pertinent question is, therefore, In what order do these two structural elements arise? That is, does sheet formation precede the formation of the associated module, thereby constraining the intervening chain and encouraging module formation, or does it follow as a result of the folding of this module? Leaving aside the very real challenge of bringing two distant parts of the chain together by diffusion alone, we raise two arguments against early β-sheet formation from our analysis. First, we find no evidence for a clear nucleation signature in the polypeptide sequence, nor do we find any interstrand pairing preferences specific to β-nucleation, making it unclear how correct strand registration could be reliably achieved. It may be argued that such signatures may yet be determined; however, we point out that we are far from the first group to examine patterning within β-strands, and thus far no specific β-nucleation signatures have emerged after many decades of research (4–9, 13–18). Second, although early β-sheet formation will certainly constrain the chain in the immediate vicinity of the sheet, this does not imply that all of the intervening residues must necessarily fold into a single structural module. Specifically, given that more than 20% of the 2-stranded β-sheets in the dataset are separated by more than 30 residues, we would expect that there would be numerous examples where the intervening chain folds into multiple structural modules, each perhaps interacting with other parts of the folding polypeptide chain, or cases where the intervening residues do not adopt a structural module at all, instead opting for long stretches of dispersed random coil. However, except for several instances that involve the

chain termini, which we discuss below, we find only one instance where the intervening residues fold into multiple, small modules (PDB code 1PBY, chain A, residues 29–31 and 112–114; Fig. 1). These two arguments together with the magnitude of the conformational search required to bring two nucleating strands together by diffusion alone, strongly suggest that β-sheet nucleation arises as a byproduct of module formation and not the other way around. Early, local folding that results in the formation of a relatively stable, compact structural module would serve to anchor the two strands involved in β-sheet nucleation near one another, dramatically reducing the conformational search required to correctly align them for β-sheet formation.

Structural anchoring of the two nucleating strands in close spatial proximity, therefore, appears to be a necessary condition for β-sheet formation. However, it does not appear to be a sufficient condition. Were it sufficient, the docking of two structural modules that brought separate coil regions in close proximity might reasonably be expected to produce an intermodule two-stranded β-sheet. As noted above, however, the only case in the dataset where the docking of two modules produces a two-stranded β-sheet is protein 1PBY (Fig. 1). We, therefore, postulate that, in addition to the rigidity provided by module formation that holds two nucleating strands in the same vicinity, considerable chain flexibility is also required to enable the two backbones to properly align themselves in close proximity for β-nucleation. The docking of two stable, folded modules presumably lacks sufficient flexibility to foster nucleation, explaining the dearth of two-stranded β-sheets composed of strands from different folding modules.

Moreover, a β-nucleation model that involves both rigidity and flexibility may now adequately account for the remaining exceptions to the single module rule for β-nucleation. In addition to the one clear exception discussed above (PDB code 1PBY, Fig. 1), there are seven other two-stranded β-sheets in the dataset that involve multiple folding modules (Table 3). Notably, all have strands that involve chain termini. Moreover, the only case of a larger β-sheet that does not contain two adjacent strands in the same folding module also involves a strand at the chain termini (PDB code 1FN9, chain A, residues 127–131 and 361–364; supplemental Fig. SF1). Chain termini are generally very flexible regions, frequently absent in protein structures determined by x-ray crystallography. If we generalize the notion of a single structural module to a single folding entity (module, domain, or entire protein chain) that provides the needed structural support to keep two regions of coil in close proximity, and we assume, quite reasonably, that these seven β-sheets involving chain termini form very late in the folding process, the basic criteria of rigidity and flexibility apply; these proteins, largely folded except for the final flexible chain termini, will anchor the two strands close together to reduce the conformational search needed for β-formation, whereas the termini bring the flexibility required to enable correct backbone interactions for β-nucleation.

This model of β-sheet formation involving both structural rigidity and chain flexibility reduces β-nucleation to a local phenomenon; either local in sequence, when there are few residues separating the two nucleating strands, or local in space, as a result of prior folding events that bring and hold the nucleating strands in close proximity. Our observations of the structural environments around two-stranded β-sheets lend support to the hydrophobic zipper model of protein folding put forth by Dill *et al.* (20), which also asserts that folding is essentially local in nature, either S-local (local in sequence) or T-local (local in topology). Moreover, a recent study postulating that sheet nucleation might occur at strand termini based on hydrophobic patterning (40) is also consistent with a "zipping" up model of β-sheet formation that proceeds from the more constrained end of the sheet toward the more flexible end. This conception of β-sheet formation is consistent with a framework model of folding wherein later events in protein folding build on earlier ones and suggests that, at least in some stages of protein folding, folding pathways do exist. Under this model, proteins solve the Levinthal paradox by topologically local interactions, where the specific topologies that emerge are based on probabilities ultimately derived from the underlying amino acid sequences.

## REFERENCES


1. Pauling, L., and Corey, R. B. (1951) *Proc. Natl. Acad. Sci. U.S.A.* **37,** 251–256
2. Pauling, L., and Corey, R. B. (1951) *Proc. Natl. Acad. Sci. U.S.A.* **37,** 729–740
3. Irvine, G. B., El-Agnaf, O. M., Shankar, G. M., and Walsh, D. M. (2008) *Mol. Med.* **14,** 451–464
4. Chou, P. Y., and Fasman, G. D. (1974) *Biochemistry* **13,** 211–222
5. Lifson, S., and Sander, C. (1979) *Nature* **282,** 109–111
6. Wouters, M. A., and Curmi, P. M. G. (1995) *Proteins* **22,** 119–131
7. Zhu, H., and Braun, W. (1999) *Protein Sci.* **8,** 326–342
8. Pal, D., and Chakrabarti, P. (2000) *Acta Crystallogr. D Biol. Crystallogr.* **56,** 589–594
9. Colloc'h, N., and Cohen, F. E. (1991) *J. Mol. Biol.* **221,** 603–613
10. Chan, A. W., Hutchinson, E. G., Harris, D., and Thornton, J. M. (1993) *Protein Sci.* **2,** 1574–1590
11. Ho, B. K., and Curmi, P. M. G. (2002) *J. Mol. Biol.* **317,** 291–308
12. Zhang, C., and Kim, S. H. (2000) *J. Mol. Biol.* **299,** 1075–1089
13. von Heijne, G., and Blomberg, C. (1977) *J. Mol. Biol.* **117,** 821–824
14. Lifson, S., and Sander, C. (1980) *J. Mol. Biol.* **139,** 627–639
15. Hutchinson, E. G., Sessions, R. B., Thornton, J. M., and Woolfson, D. N. (1998) *Protein Sci.* **7,** 2287–2300
16. Cootes, A. P., Curmi, P. M., Cunningham, R., Donnelly, C., and Torda, A. E. (1998) *Proteins* **32,** 175–189
17. Mandel-Gutfreund, Y., Zaremba, S. M., and Gregoret, L. M. (2001) *J. Mol. Biol.* **305,** 1145–1159
18. Fooks, H. M., Martin, A. C., Woolfson, D. N., Sessions, R. B., and Hutchinson, E. G. (2006) *J. Mol. Biol.* **356,** 32–44
19. Muñoz, V., Thompson, P. A., Hofrichter, J., and Eaton, W. A. (1997) *Nature* **390,** 196–199
20. Dill, K. A., Fiebig, K. M., and Chan, H. S. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90,** 1942–1946
21. Ramírez-Alvarado, M., Serrano, L., and Blanco, F. J. (1997) *Protein Sci.* **6,** 162–174
22. Viguera, A. R., Jiménez, M. A., Rico, M., and Serrano, L. (1996) *J. Mol. Biol.* **255,** 507–521
23. Blanco, F., Ramírez-Alvarado, M., and Serrano, L. (1998) *Curr. Opin. Struct. Biol.* **8,** 107–111
24. Lacroix, E., Kortemme, T., Lopez de la Paz, M., and Serrano, L. (1999) *Curr. Opin. Struct. Biol.* **9,** 487–493
25. Hughes, R. M., and Waters, M. L. (2006) *Curr. Opin. Struct. Biol.* **16,** 514–524
26. Syud, F. A., Stanger, H. E., and Gellman, S. H. (2001) *J. Am. Chem. Soc.* **123,** 8667–8677

27. Jäger, M., Nguyen, H., Crane, J. C., Kelly, J. W., and Gruebele, M. (2001) *J. Mol. Biol.* **311,** 373–393
28. Espinosa, J. F., Syud, F. A., and Gellman, S. H. (2002) *Protein Sci.* **11,** 1492–1505
29. Santiveri, C. M., Santoro, J., Rico, M., and Jiménez, M. A. (2004) *Protein Sci.* **13,** 1134–1147
30. Dyer, R. B., Maness, S. J., Peterson, E. S., Franzen, S., Fesinmeyer, R. M., and Andersen, N. H. (2004) *Biochemistry* **43,** 11560–11566
31. Kuo, N. N., Huang, J. J., Miksovska, J., Chen, R. P., Larsen, R. W., and Chan, S. I. (2005) *J. Am. Chem. Soc.* **127,** 16945–16954
32. Petrovich, M., Jonsson, A. L., Ferguson, N., Daggett, V., and Fersht, A. R. (2006) *J. Mol. Biol.* **360,** 865–881
33. Deechongkit, S., Nguyen, H., Jager, M., Powers, E. T., Gruebele, M., and Kelly, J. W. (2006) *Curr. Opin. Struct. Biol.* **16,** 94–101
34. Muñoz, V., Henry, E. R., Hofrichter, J., and Eaton, W. A. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95,** 5872–5879
35. Sung, S. S. (1999) *Biophys. J.* **76,** 164–175
36. Dinner, A. R., Lazaridis, T., and Karplus, M. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96,** 9068–9073
37. Colombo, G., De Mori, G. M., and Roccatano, D. (2003) *Protein Sci.* **12,** 538–550
38. Santiveri, C. M., Jiménez, M. A., Rico, M., Van Gunsteren, W. F., and Daura, X. (2004) *J. Pept. Sci.* **10,** 546–565
39. Noguchi, T., Matsuda, H., and Akiyama, Y. (2001) *Nucleic Acids Res.* **29,** 219–220
40. Kabsch, W., and Sander, C. (1983) *Biopolymers* **22,** 2577–2637
41. Pickford, A. R., and Campbell, I. D. (2004) *Chem. Rev.* **104,** 3557–3566
42. Wathen, B., and Jia, Z. (2009) *Int. J. Mol. Sci.* **10,** 1567–1589