The N-terminal to C-terminal motif in protein folding and function

Mallela M. G. Krishna* and S. Walter Englander

Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6059

Contributed by S. Walter Englander, December 15, 2004

Essentially all proteins known to fold kinetically in a two-state manner have their N- and C-terminal secondary structural elements in contact, and the terminal elements often dock as part of the experimentally measurable initial folding step. Conversely, all N–C no-contact proteins studied so far fold by non-two-state kinetics. By comparison, about half of the single domain proteins in the Protein Data Bank have their N- and C-terminal elements in contact, more than expected on a random probability basis but not nearly enough to account for the bias in protein folding. Possible reasons for this bias relate to the mechanisms for initial protein folding, native state stability, and final turnover.

loop closure | terminal contacts | stability | turnover

Although simple physical principles provide no apparent reason why proteins should bring their far N and C termini into contact (1), even a cursory examination of known structures shows that many proteins do so. In earlier work, Thornton and Sibanda (2) and Christopher and Baldwin (3) looked for some statistical tendency for N- and C-terminal amino acid residues to be in close proximity. By using the limited database of 72 structures available at the time, these workers found no statistically significant tendency for the far-terminal residues to approach each other but a significant preference emerged when longer terminal segments were considered [10 residues, proximity within 10 Å Cα to Cα (2), or 6 residues endowed with complete flexibility (3)].

Recent results on protein folding that suggest a special role for terminal secondary structure elements encouraged us to reexamine the issue. We redefine the question in terms of overt contacts ($\leq$5 Å) between nonhydrogen atoms in terminal secondary structural elements, and from this point of view consider the large body of protein folding literature and the now greatly expanded Protein Data Bank (PDB) (4).

Methods

The list of two-state folding proteins in Table 1, which is published as supporting information on the PNAS web site, was compiled from prior listings in the literature (5–10). A representative protein data set from the PDB was selected. Although several methods are available to cull protein structures according to their similarity in sequence and structure, we used PDB-REPRDB web server (11) (http://mbs.cbrc.jp/pdbreprdb-cgi/reprdb_menu.pl; March 17, 2004 update) because of the ease of selecting protein chains based on different criteria and the provision of corresponding SCOP (12) codes for each protein chain. A nonredundant protein set (no membrane proteins; minimum chain length, 40; x-ray resolution, $\leq$2 Å, *R*-factor, $\leq$0.3; all NMR structures; 2,120 chains) was selected at 30% sequence similarity and 10-Å structure similarity (rms deviation between Cα atoms).

Protein chains were classified into single, multi, and unknown domains based on individual SCOP codes. PDB files were downloaded by using the PERL script getPdbStructures.pl (http://msdlocal.ebi.ac.uk/docs/rcsb/pdb/software/getPdbStructures.html). Secondary structures (α-helix or β-strand) were assigned by using DSSP (13) (www.cmbi.kun.nl/gv/dssp). The minimum lengths of an α-helix and a β-strand are 4 and 2, respectively. It should be noted that the DSSP helix length is less by one amino acid on both ends than International Union of Pure and Applied Chemistry–International Union of Biochemistry recommendations. Other DSSP structures including 3_{10}, π-helix, and isolated β-bridges, -turns, and -bends were omitted in our analysis.

All structure analysis computer programs were written in ANSI C. Because the residue numbering in PDB structure files is not sequential, each amino acid has three identifiers in our data analysis programs: PDB author-assigned residue number, insertion code, and DSSP assigned sequential number. Each step in the programs was initially checked with dummy data sets. Histograms were generated by using SIGMAPLOT 2001.

To test the possibility of higher N-element to C-element (N–C) contact probability in protein fragments, we omitted proteins that are fragments of longer proteins by scanning the COMPND variable for the keyword "fragment" in PDB files. Some of the later sequence culling was done by using the PISCES (14) web server (www.fccc.edu/research/labs/dunbrack/pisces) because PDB-REPRDB does not allow culling based on user-selected sequences.

Results

The N–C Secondary Structural Motif in Protein Folding. Native state hydrogen exchange studies indicate that protein folding may often use secondary structural elements rather than individual amino acids as building blocks for folding (foldons) (15–17). For example, cytochrome *c* folds by assembling five native-like foldon units in a stepwise manner to progressively construct the native protein (18–21). Similar experiments with some other proteins find similar results (15–17, 22–27). Therefore, we focus the present analysis on secondary structural elements. Related observations focus attention especially on the terminal secondary structural elements. For example, hydrogen exchange pulse labeling (28, 29) and mutational (30) experiments show that the N- and C-terminal helices of cytochrome *c* dock to form an initial kinetic folding intermediate. In fact, these helices appear to form and dock in the initial folding transition state (31, 32).

The folding literature indicates that many proteins dock their terminal elements as a first step in folding. In apomyoglobin, the N-terminal A helix and the C-terminal GH bihelix dock in an initial folding intermediate (33–35). Studies of folding transition states by ϕ-analysis point to key interacting residues that have high ϕ-values ($\geq$0.35) in the N- and C-terminal secondary structural elements of a number of proteins. Examples include the terminal α-helices of acyl-coA binding protein (36), spectrin R16 (37), and the bacterial immunity proteins Im7 and Im9 (38, 39); the terminal β-strands of muscle acylphosphatase (40), human procarboxypeptidase A2 (40, 41), and PsaE (42); the terminal β-hairpins of protein G (43), FN_III domain 10 (44), and CspB (45, 46); and the N-terminal α-helix and C-terminal

BIOPHYSICS

Abbreviations: PDB, Protein Data Bank; N–C, N-element to C-element.

*To whom correspondence should be addressed. E-mail: mmg@hx2.med.upenn.edu.

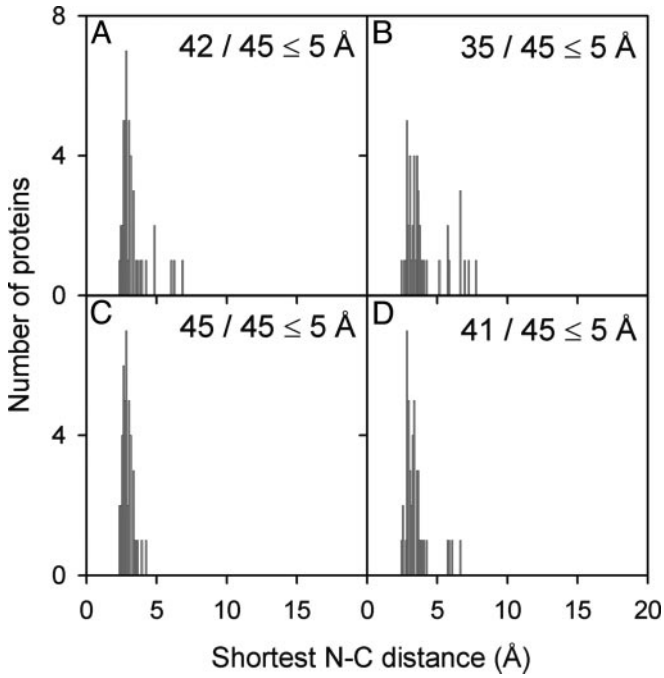

Fig. 1. Distance between N- and C-terminal secondary structural elements in 45 two-state folding proteins (listed in Table 1). The distances refer to nonhydrogen atoms in one (*A* and *C*) or two nonoverlapping (*B* and *D*) pairs of contacting residues. In *A* and *B*, α-helix and β-strand were considered as structural elements. In *C* and *D*, a terminal β-hairpin, if present, was taken as a single structural element.

β-hairpin of CI2 (47). The recently developed ψ-analysis of folding transition states identified contacting residue pairs in the N- and C-terminal β-strands of ubiquitin (48, 49). In the dimeric proteins Trp repressor (50) and Arc repressor (51), interaction of terminal secondary structural elements between monomers appears to be a key event in folding.

A few other proteins do not appear to form their N–C contacts first, including cytochrome b562 (24, 26), and some Src homology 3 (SH3) domains [src (52), α-spectrin (53), fyn (54)]. The C-terminal α-helix of the sso7d SH3 domain (55) has high ϕ values, but not the N-terminal β-strand. However, molecular dynamics simulations place the terminal elements close together in the majority of the transition state ensembles of src, α-spectrin, and fyn SH3 domains (56). Also in PsaE, a structural analog of SH3 domains, residues in both terminal β-strands have high ϕ values (42).

Other observations further emphasize the N–C contact motif. Table 1 lists proteins that are known to fold to their native state in a kinetically two-state manner, i.e., without the obvious accumulation of intermediates. We limit the list to the 45 proteins that have two or more secondary structural elements and a minimum of 40 residues to minimize bias in the analysis for N–C contacts due to an overemphasis on small proteins. It can be noted that one-quarter of the two-state folding proteins listed in Table 1 are longer than 100 residues and they extend out to VlsE, which has 341 residues.

Fig. 1 shows the distribution of shortest N–C distances (α-helix or β-strand) computed for these proteins. If we tally a contact when nonhydrogen atoms in two residues come within 5 Å, 93% of the two-state proteins have at least one pair of N-terminal to C-terminal residue–residue contacts (Fig. 1*A*) and 78% have at least two nonoverlapping pairs (Fig. 1*B*). If terminal β-hairpins are considered single structural elements, the probability rises to 100% for one contacting residue pair (Fig. 1*C*) and 91% for two nonoverlapping pairs of contacting residues (Fig. 1*D*). Decreasing the cutoff to 4 Å barely changes these numbers.

In summary, available folding studies seem to point to some special role for terminal element interactions in kinetic protein folding. However, the possibility exists that these results may, in fact, represent some unexpected character of proteins more generally. Therefore, we considered the probability of terminal contacts in the total PDB. Analysis (see below) shows that the bias toward an N–C motif in the PDB is much more common than might be expected on a random basis, but not nearly so striking as in the folding literature.

The Structural Data Set. The PDB was culled at 30% sequence similarity and 10-Å structure similarity (PDB-REPRDB) so that a closely related family is not overrepresented by many members. Fig. 2 characterizes the number, length, placement, and contacts of the secondary structural elements (α-helix/β-strand) in 1,559 single SCOP domains.

The number of secondary structural elements per protein in our data set peak at 5 or 6 but extend out to very many more (arithmetic mean at 12) (Fig. 2*A*). Helices tend to be longer than β-strands, with modal lengths of 10–11 and 4–5, respectively, and skewing to greater lengths, especially for helices (Fig. 2 *B* and *C*).

Terminal elements show a similar length distribution (Fig. 2 *B* and *C*). There is a modest preference for β-strand over α-helix in N-terminal elements (1.4 times) and for α-helix in the C-terminal elements (1.3 times; Fig. 2*D*). This disparity is in qualitative agreement with the earlier observations of Thornton and Chakauya (57) on a much smaller data set (54 nonhomologous proteins). N- and C-terminal elements are very similar in terms of their length and number of long-range contacts (Fig. 2 *E* and *F*).

The Probability of Terminal Contacts. Fig. 3 shows the distribution within the database of shortest distances between N- and C-terminal structural elements (α-helix or β-strand). Distances $\leq$5 Å represent overt contacts. Half of the proteins in the database bring their N- and C-terminal elements into direct contact (Fig. 3*A*). When the contact criterion is extended to at least two residues in each terminal element, the contact probability falls to 37% (Fig. 3*B*). These numbers are relatively insensitive to different selection and culling criteria (Table 2, which is published as supporting information on the PNAS web site).

When terminal β-strands occur as β-hairpins (41% of cases), it seems reasonable to consider the entire hairpin as a single structural element. On this basis, the N–C contact probability is 54% for one or more and 44% for two or more nonoverlapping pairs of contacting residues (Fig. 3 *C* and *D*).

Selective interaction is not seen for terminal chain lengths more distal than the terminal secondary structure elements. Their contact probability is about half that of the terminal secondary elements, close to the random probability (see below), whereas the far-terminal residues themselves have much lower contact probability ($\approx$1%). The positive results found by Thornton and Sibanda (2) and by Christopher and Baldwin (3) for longer terminal segments are due to incursion into the secondary structural elements and/or to the more flexible criterion for proximity.

Effect of Chain Length. Any given terminal element has a high probability of contacting its near-neighbor elements (77% experience nearest neighbor contact and 53% next-neighbor contact). To minimize near neighbor bias, we eliminated proteins that are shorter than 60 residues, that have fewer than four secondary structural elements, or that have terminal elements separated by $<$30 residues. These constraints reduce our data set to 1,363 proteins, but they do not significantly change the computed N–C contact probability (46% for $\geq$1 contact and 33% for $\geq$2 nonoverlapping contacting pairs) (Table 2).

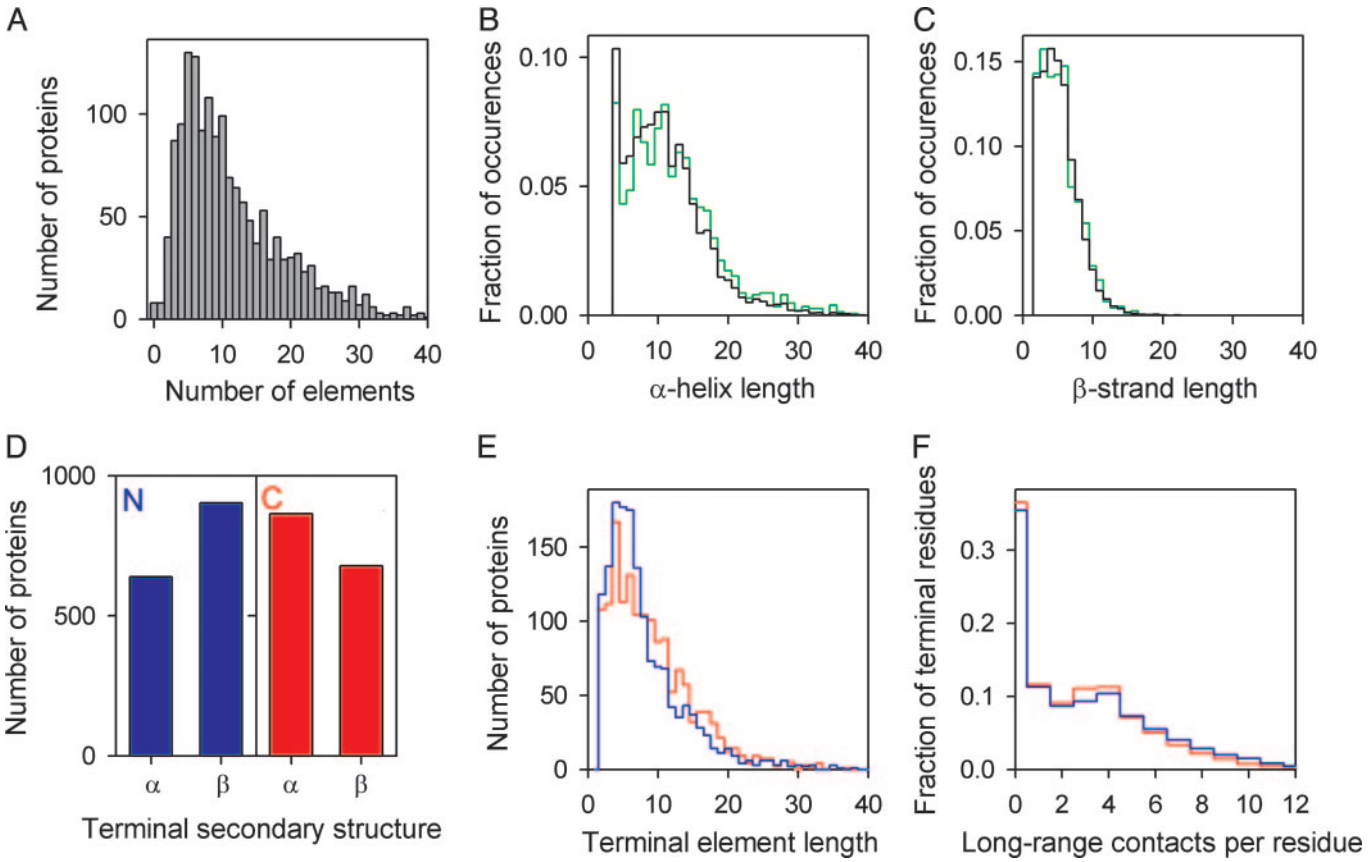

Fig. 2. Characterization of the PDB culled at 30% sequence similarity and 10-Å structure similarity by using PDB-REPRDB. (*A–C*) The population distribution of secondary structural elements (α-helix/β-strand) and their lengths in 1,559 single SCOP domains. (*B* and *C*) Black and green represent all elements and terminal elements respectively. (*D–F*) The terminal secondary structure preference, the terminal element length distribution, and the total number of long-range contacts (>30 intervening residues) per terminal element residue (in 1,543 single SCOP domains with at least two secondary structural elements). Blue and red represent N and C elements, respectively.

Fig. 4 shows how the probability of N–C contact varies with the length of the protein. As the number of intervening elements increases to large values, the probability of N–C contact falls but only to ≈25%, which compares with ≈13% for terminal element to middle element contacts (measure of random probability, see below and Tables 2 and 3, which are published as supporting information on the PNAS web site).

N–C Contact Proteins Versus N–C No-Contact Proteins. To gauge the statistical significance of these observations, we compared the probability that terminal elements contact each other with the probability that they contact any other element. Fig. 5 shows results for a set of single domain proteins that have twelve secondary structural elements, the mean number per protein in the data set. The different panels show the probability distribution for all proteins and for the subsets that have no N–C contacts and one or more N–C contacts.

For N–C no-contact proteins (Fig. 5*A*), the probability that a terminal element contacts another element decreases as the number of intervening elements increases. This is the expected result on a simple physical basis.

For the N–C contact proteins (Fig. 5*B*), a different pattern emerges. The probability that a terminal element contacts another element decreases as the element separation increases, but it increases again as the other terminus is approached. The falling pattern on the left side of the figure is caused by the neighbor effect and its decrease as sequence distance increases. This same pattern, sloping from the other far-terminal element, is similarly selected for in the population used to construct Fig. 5*B* (N–C contact proteins). However, when the total population is considered (N–C contact plus N–C no-contact; Fig. 5*C*), the smile pattern continues to be seen, demonstrating the dominant tendency of proteins to bring N- and C-elements together.

The smile pattern is present independently of the number of secondary elements in the protein set examined and also with two contacting residues as the criterion. When progressively larger proteins are examined, both the terminal contact probability and the terminal to middle contact probability decrease, but the terminal to terminal probability is always about double the terminal to middle probability.

Terminal Contact Probability Versus Random Probability. The smile pattern observed in native proteins (Fig. 5) is not a characteristic of random-flight protein chains. In an unfolded random-flight chain, the spatial distance between any two residues continuously increases with the number of intervening residues (2). For hypothetical random globular proteins generated by a Monte Carlo simulation with the chain constrained within an ellipsoid that matches the protein packing density, the segment contact probability decreases as sequence separation increases and then reaches a plateau level that is close to the minimum values observed for native proteins (see figure 6 of ref. 2 and Fig. 5). This nonbiased level can be taken to represent the random contact probability.

Over the entire data set, the probability observed for N–C contact is 2.3 times the random probability (terminal to middle element contact) for one or more pairs of contacting residues

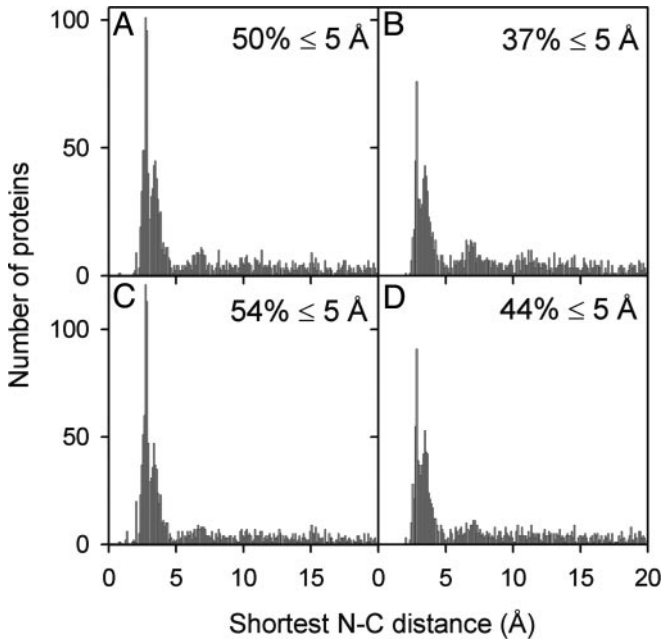

Fig. 3. Distance distribution for N–C contacts for single domain proteins in the overall PDB (compare Fig. 1). (*A* and *C*) Shortest residue–residue contact distances. (*B* and *D*) Shortest contact distances for a second pair of nonoverlapping contacting residues. In *A* and *B*, α-helix and β-strand were considered as structural elements. In *C* and *D*, a terminal β-hairpin, if present, was taken as a single structural element.

and it is 3 times for two or more nonoverlapping contacting pairs. These ratios remain nearly the same even when we consider larger proteins with at least 13 secondary structural elements and 160 residues, for which the N–C contact probability reaches a plateau level with increasing protein length (Fig. 4 and Tables 2 and 3). The N–C contact probability is much lower in the case of multidomain proteins (15% for one and 7% for two nonoverlapping contacting pairs), which provides another measure of the random probability.

These and other results, gathered in Table 3, confirm that the terminal contact probability in the overall PDB is much larger than that determined by random probability.

The Protein Fragment Problem. Nearly one-fourth of single-domain proteins in the PDB are protein fragments. Perhaps such fragments have an enhanced chance for their terminal elements to be in contact (58). We tested a data set of single domain proteins

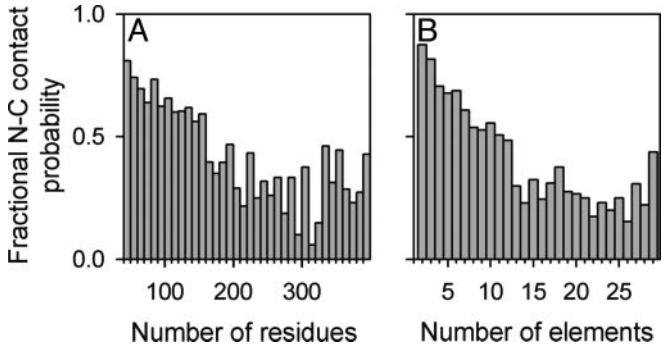

Fig. 4. N–C contact probability as a function of protein length. The data set is for 1,543 single domain proteins that have at least two secondary structural elements. Here, α-helix and β-strand were considered as single structural elements.

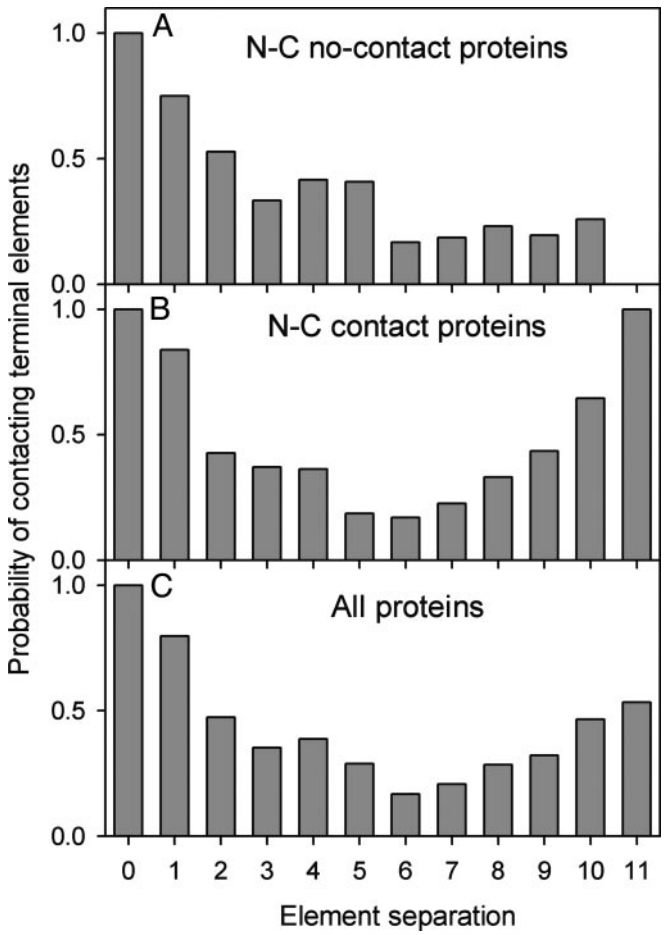

Fig. 5. Probability of contact between terminal elements and any other element as a function of element separation. The distributions shown are for proteins that have 12 secondary structural elements (α-helix plus β-strand). Results are for the N–C no-contact proteins (*A*), for N–C contact proteins with $\geq$1 contact (*B*), and for the summed data set (*C*). To minimize noise, the bar height for element separation equal to one averages contacts between the N-terminal and N + 1 elements and between the C-terminal and C − 1 elements, and similarly for larger separations.

that are not fragments of larger proteins. From all of the available protein chains (PDB-REPRDB; 12,316 chains), we excluded all multi and unknown domains by using SCOP codes and also single domains that are fragments of larger proteins. We culled at 30% sequence similarity by using PISCES (14) and eliminated proteins with <60 residues, four secondary elements, or terminal elements separated by <30 residues.

In the resulting 964 nonfragment single domain proteins, the N–C contact probability is ≈43% for ≥1 contact and 31% for ≥2 nonoverlapping contacting pairs, much the same as for all (fragment plus nonfragment) proteins (Table 2). Thus, the inclusion of protein fragments in the data set does not significantly bias toward higher N–C contact probability.

Are Two-State Folders Different? The N–C contact frequency found for two-state folders is larger than the frequency found for proteins in the PDB. Is this difference statistically significant? One can ask: if 45 proteins are drawn randomly from the PDB, with the same size as the two-state proteins in Fig. 1*A* (and Table 1), what is the probability that 42 or more will happen to exhibit an N–C contact? The answer is that this result will occur by chance in 40 per million trials ($P = 4 \times 10^{-5}$; calculated from Fig. 4*A*). The same question, asked by

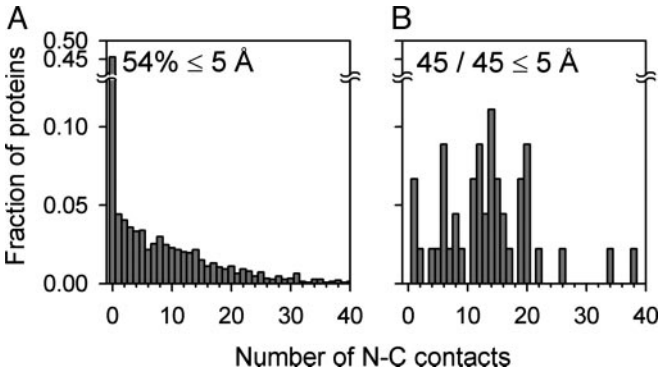

Fig. 6. Population distribution of all N–C contacts (overlapping plus nonoverlapping residue pairs). (*A*) For 1,535 single SCOP domains with at least two secondary structural elements. (*B*) For the 45 two-state proteins. In addition to α-helix and β-strand, terminal β-hairpins were taken as single structural elements.

using a different criterion (Fig. 1*C*; β-hairpin as single structural element), returns a much lower probability ($P = 6 \times 10^{-7}$). These probabilities do not change significantly when we exclude two-state proteins that have >30% sequence similarity. In short, the high terminal element contact frequency found for two-state folders is a selected property, not accounted for by the N–C contact probability in the parent PDB.

How extensive are the N–C contacts? Among proteins in the PDB (Fig. 6*A*), 46% make no N–C contacts and the ones that do make contacts show no apparent preference for contact extent. Two-state folders insist on N–C contact and appear to prefer sizeable contacts in the range of 5 to 20 residue–residue pairs but smaller and larger interactions occur (Fig. 6*B* and Fig. 7, which is published as supporting information on the PNAS web site; counting both overlapping and nonoverlapping residue pairs).

Discussion

We find that the N–C motif is present in close to 100% of known two-state folding proteins. The N–C "background" level in the overall PDB is sizeable, but it does not account for the extremely high N–C contact level seen for the two-state folders. In fact, even some of the N–C background in the PDB seems likely to be due to its selection for protein folding purposes.

The N–C contact tends to form as part of an initial step in folding. This bias is not expected on simple physical grounds, which favor the initial interaction and zipping up of local sites and militate against the kinetic docking of distant elements (1). Homopolymers condense by bringing together near neighbor units rather than sequentially distant regions (59). Randomly generated globular structures show no bias toward far terminal interactions (2). Known differences between amino acids in terminal and nonterminal protein regions (60–62) do not seem adapted to selectively favor N–C contacts.

These observations suggest that the N–C motif has been evolutionarily selected for some functional advantage and is built into the structural design of many proteins. Apparent possibilities relate to folding at the beginning of the protein life cycle, and to native state stability and turnover at the end.

Protein Folding. The physical basis for two-state folding is the fact that folding is rate-limited by an initial barrier B1 (17, 63) before which no stable intermediate is formed (64); otherwise, folding would not appear to be a two-state process. Essentially by definition, B1 represents a time-requiring free-energy-uphill conformational search for a transition state that can allow protein chains to begin to go forward in a free-energy-downhill manner. We interpreted experimental results to indicate that successful folding transition states (B1) consist of the initially collapsed protein chain pinned into a native-like chain folding topology (63). This view seems to be confirmed by the success of the contact order formulation for two-state proteins discovered by Plaxco and coworkers (8, 65), and similar formulations developed by others, all of which assume that productive folding begins by finding the correct topology in the functional (initial) rate-limiting step.

An apparent functional rationale for two-state folding, with the initial barrier being rate-limiting, is that it avoids the prolonged occupation of collapsed partially folded states that would expose proteins to unwanted intermolecular aggregation and proteolysis. This is desirable both during the initial folding process and subsequently during the life of the protein because native proteins repeatedly unfold and refold even under native conditions (18). To promote formation of a "correct" native topology initially and to avoid later fraying-dependent proteolysis and aggregation, it seems useful to correctly orient and tie down the chain ends in the initial folding-collapse step, keep them securely tied down in the native condition and in transient intermediates that form during folding and unfolding, and allow their release only in the final re-unfolding step.

One can then ask whether N–C no-contact proteins tend to fold in a non-two-state manner. Few N–C no-contact proteins have been studied, due perhaps to the selection bias in folding studies toward smaller proteins, which tend to have N–C contacts (Fig. 4). A survey shows that 11 of the 70 proteins so far studied have no N–C contact (Tables 1 and 4, which are published as supporting information on the PNAS web site, and refs. 6 and 66). All 11 fold by non-two-state kinetics (Table 4). As before, this biased result would occur by chance selection from the PDB background with low probability ($P = 4 \times 10^{-4}$; calculated from Fig. 4*A*).

In considering these issues, it should be noted that the distinction between two-state and non-two-state folders can be confused by *in vitro* experimental factors. A number of N–C contact proteins, when examined fairly, in fact fold heterogeneously with two-state and non-two-state subpopulations. Heterogeneity and non-two-state folding are due to later barriers (B2 > B1) which, we believe (16, 17, 67), are not intrinsic to the folding process but reflect the probabilistic insertion ($0 \leq P \leq 1$) of misfolding-dependent errors, often as an artifact of *in vitro* experimental conditions (proline misisomerization, aggregation, incorrect disulfides, heme misligation, etc.). In addition, non-two-state folding of N–C no-contact proteins may be due to second barriers (after B1) that are really intrinsic to the folding process; however, these barriers remain to be experimentally identified. How much of this behavior occurs *in vivo* where it might exert evolutionary pressure toward N–C contact topologies remains to be seen.

Other Functions: Protein Stability and Turnover. It seems interesting that terminal interactions in general, even without N–C contact, may play some special role in protein stability and turnover. A number of proteins are known to become destabilized and even unfold when a terminal length is removed. Examples include cytochrome *c* (68), ribonuclease A (69, 70), staphylococcal nuclease (71), CI2 (72), Titin (73), TNfn3 (74), bovine pancreatic DNase (75), botulinum neurotoxin type A light chain (76), and Fyn SH3 (K. Plaxco, personal communication). Protein turnover is controlled by mechanisms that unfold proteins preparatory to proteolytic destruction, perhaps by forcefully pulling out a terminal segment (77–82).

We thank Robert Baldwin, David Goldenberg, Neville Kallenbach, Leland Mayne, Marcos Milla, Kevin Plaxco, George Rose, Tobin Sosnick, Janet Thornton, and Joshua Wand for helpful discussions and comments on the manuscript. This work was supported by National Institutes of Health Grant GM31847.

1. Weikl, T. R., Palassini, M. & Dill, K. A. (2004) *Protein Sci.* **13,** 822–829.
2. Thornton, J. M. & Sibanda, B. L. (1983) *J. Mol. Biol.* **167,** 443–460.
3. Christopher, J. A. & Baldwin, T. O. (1996) *J. Mol. Biol.* **257,** 175–187.
4. Berman, H. M., Westbrook, J., Z. Feng, Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
5. Bai, Y., Zhou, H. & Zhou, Y. (2004) *Protein Sci.* **13,** 1173–1181.
6. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N. & Finkelstein, A. V. (2003) *Proteins Struct. Funct. Genet.* **51,** 162–166.
7. Jackson, S. E. (1998) *Fold. Des.* **3,** R81–R91.
8. Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. (2000) *Biochemistry* **39,** 11177–11183.
9. Sato, S. & Raleigh, D. P. (2002) *J. Mol. Biol.* **318,** 571–582.
10. Jones, K. & Wittung-Stafshede, P. (2003) *J. Am. Chem. Soc.* **125,** 9606–9607.
11. Noguchi, T. & Akiyama, Y. (2003) *Nucleic Acids Res.* **31,** 492–493.
12. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. & Murzin, A. G. (2004) *Nucleic Acids Res.* **32,** D226–D229.
13. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
14. Wang, G. & Dunbrack, R. L., Jr. (2003) *Bioinformatics* **19,** 1589–1591.
15. Chamberlain, A. K. & Marqusee, S. (2000) *Adv. Protein Chem.* **53,** 283–328.
16. Englander, S. W. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29,** 213–238.
17. Krishna, M. M. G., Lin, Y. & Englander, S. W. (2004) *J. Mol. Biol.* **343,** 1095–1109.
18. Bai, Y., Sosnick, T. R., Mayne, L. & Englander, S. W. (1995) *Science* **269,** 192–197.
19. Rumbley, J., Hoang, L., Mayne, L. & Englander, S. W. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 105–112.
20. Krishna, M. M. G., Lin, Y., Rumbley, J. N. & Englander, S. W. (2003) *J. Mol. Biol.* **331,** 29–36.
21. Maity, H., Maity, M. & Englander, S. W. (2004) *J. Mol. Biol.* **343,** 223–233.
22. Raschke, T. M. & Marqusee, S. (1998) *Curr. Opin. Biotechnol.* **9,** 80–86.
23. Fuentes, E. J. & Wand, A. J. (1998) *Biochemistry* **37,** 9877–9883.
24. Fuentes, E. J. & Wand, A. J. (1998) *Biochemistry.* **37,** 3687–3698.
25. Yan, S., Kennedy, S. D. & Koide, S. (2002) *J. Mol. Biol.* **323,** 363–375.
26. Chu, R., Pei, W., Takei, J. & Bai, Y. (2002) *Biochemistry* **41,** 7998–8003.
27. Silverman, J. A. & Harbury, P. B. (2002) *J. Mol. Biol.* **324,** 1031–1040.
28. Roder, H., Elöve, G. A. & Englander, S. W. (1988) *Nature* **335,** 700–704.
29. Krishna, M. M. G., Lin, Y., Mayne, L. & Englander, S. W. (2003) *J. Mol. Biol.* **334,** 501–513.
30. Colón, W., Elöve, G. A., Wakem, P., Sherman, F. & Roder, H. (1996) *Biochemistry* **35,** 5538–5549.
31. Krantz, B. A., Moran, L. B., Kentsis, A. & Sosnick, T. R. (2000) *Nat. Struct. Biol.* **7,** 62–71.
32. Krantz, B. A., Srivastava, A. K., Nauli, S., Baker, D., Sauer, R. T. & Sosnick, T. R. (2002) *Nat. Struct. Biol.* **9,** 458–463.
33. Hughson, F. M., Wright, P. E. & Baldwin, R. L. (1990) *Science* **249,** 1544–1548.
34. Jennings, P. A. & Wright, P. E. (1993) *Science* **262,** 892–896.
35. Kay, M. S., Ramos, C. H. & Baldwin, R. L. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2007–2012.
36. Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiødt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. M. (1999) *Nat. Struct. Biol.* **6,** 594–601.
37. Scott, K. A., Randles, L. G. & Clarke, J. (2004) *J. Mol. Biol.* **344,** 207–221.
38. Capaldi, A. P., Kleanthous, C. & Radford, S. E. (2002) *Nat. Struct. Biol.* **9,** 209–216.
39. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003) *J. Mol. Biol.* **326,** 293–305.
40. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999) *Nat. Struct. Biol.* **6,** 1005–1009.
41. Villegas, V., Martinez, J. C., Avilés, F. X. & Serrano, L. (1998) *J. Mol. Biol.* **283,** 1027–1036.
42. Grantcharova, V., Alm, E. J., Baker, D. & Horwich, A. L. (2001) *Curr. Opin. Struct. Biol.* **11,** 70–82.
43. McCallister, E. L., Alm, E. & Baker, D. (2000) *Nat. Struct. Biol.* **7,** 669–673.
44. Cota, E., Steward, A., Fowler, S. B. & Clarke, J. (2001) *J. Mol. Biol.* **305,** 1185–1194.
45. Perl, D., Holtermann, G. & Schmid, F. X. (2001) *Biochemistry* **40,** 15501–15511.
46. Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. (2004) *J. Mol. Biol.* **339,** 555–569.
47. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 260–288.
48. Krantz, B. A., Dothager, R. S. & Sosnick, T. R. (2004) *J. Mol. Biol.* **337,** 463–475.
49. Sosnick, T. R., Dothager, R. S. & Krantz, B. A. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17377–17382.
50. Tasayco, M. L. & Carey, J. (1992) *Science* **255,** 594–597.
51. Milla, M. E. & Sauer, R. T. (1995) *Biochemistry* **34,** 3344–3351.
52. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999) *Nat. Struct. Biol.* **6,** 1016–1024.
53. Martinez, J. C. & Serrano, L. (1999) *Nat. Struct. Biol.* **6,** 1010–1016.
54. Northey, J. G. B., Nardo, A. A. D. & Davidson, A. R. (2002) *Nat. Struct. Biol.* **9,** 126–130.
55. Guerois, R. & Serrano, L. (2000) *J. Mol. Biol.* **304,** 967–982.
56. Lindorff-Larsen, K., Vendruscolo, M., Paci, E. & Dobson, C. M. (2004) *Nat. Struct. Mol. Biol.* **11,** 443–449.
57. Thornton, J. M. & Chakauya, B. L. (1982) *Nature* **298,** 296–297.
58. Aroul-Selvam, R., Hubbard, T. & Sasidharan, R. (2004) *J. Mol. Biol.* **338,** 633–641.
59. Chan, H. S. & Dill, K. A. (1991) *Annu. Rev. Biophys. Biophys. Chem.* **20,** 447–490.
60. Berezovsky, I. N., Kilosanidze, G. T., Tumanyan, V. G. & Kisselev, L. L. (1999) *Protein Eng.* **12,** 23–30.
61. Pal, D. & Chakrabarti, P. (2000) *Biopolymers* **53,** 467–475.
62. Bhattacharyya, R., Pal, D. & Chakrabarti, P. (2002) *Acta Crystallogr. D* **58,** 1793–1802.
63. Sosnick, T. R., Mayne, L. & Englander, S. W. (1996) *Proteins Struct. Funct. Genet.* **24,** 413–426.
64. Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W. & Sosnick, T. R. (2002) *J. Mol. Biol.* **324,** 359–371.
65. Plaxco, K. W., Simons, K. T. & Baker, D. (1998) *J. Mol. Biol.* **277,** 985–994.
66. Kamagata, K., Arai, M. & Kuwajima, K. (2004) *J. Mol. Biol.* **339,** 951–965.
67. Sosnick, T. R., Mayne, L., Hiller, R. & Englander, S. W. (1994) *Nat. Struct. Biol.* **1,** 149–156.
68. Sosnick, T. R., Shtilerman, M. D., Mayne, L. & Englander, S. W. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 8545–8550.
69. Anfinsen, C. B. (1956) *J. Biol. Chem.* **221,** 405–412.
70. Taniuchi, H. (1970) *J. Biol. Chem.* **245,** 5459–5468.
71. Taniuchi, H. & Anfinsen, C. B. (1969) *J. Biol. Chem.* **244,** 3864–3875.
72. de Prat Gay, G., Ruiz-Sanz, J., Neira, J. L., Corrales, F. J., Otzen, D. E., Ladurner, A. G. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 968–979.
73. Pfuhl, M., Improta, S., Politou, A. S. & Pastore, A. (1997) *J. Mol. Biol.* **265,** 242–256.
74. Hamill, S. J., Meekhof, A. E. & Clarke, J. (1998) *Biochemistry* **37,** 8071–8079.
75. Chen, W.-J., Huang, P.-T., Liu, J. & Liao, T.-H. (2004) *Biochemistry* **43,** 10653–10663.
76. Baldwin, M. R., Bradshaw, M., Johnson, E. A. & Barbieri, J. T. (2004) *Protein Expr. Pur.* **37,** 187–195.
77. Keiler, K. C., Waller, P. R. H. & Sauer, R. T. (1996) *Science* **271,** 990–993.
78. Gottesman, S., Roche, E., Zhou, Y. & Sauer, R. T. (1998) *Genes Dev.* **12,** 1338–1347.
79. Flynn, J. M., Neher, S. B., Kim, Y.-I., Sauer, R. T. & Baker, T. A. (2003) *Mol. Cell* **11,** 671–683.
80. Glickman, M. H. & Ciechanover, A. (2002) *Physiol. Rev.* **82,** 373–428.
81. Varshavsky, A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 12142–12149.
82. Hochstrasser, M. (1996) *Annu. Rev. Genet.* **30,** 405–439.

Table 1. List of all two-state proteins (^{3} 2 elements and 40 amino acids; not culled for similarity)

Protein	PDB code	Length
E2P	2PDD_	43
Engrailed Homeodomain	1ENH_	54
Protein G	1PGB_	56
NTL9	1CQUA	56
Protein A B-domain	1BDC_	60
α-spectrin SH3 domain	1SHG_	62
Sso7d SH3 domain	1BF4A	63
Src SH3 domain	1SRM_	64
CI2	2CI2I	65
CspB (B. caldolyticus)	1C9OA	66
CspB (T. maritime)	1G6PA	66
Fyn-SH3 domain	1NYF_	67
CspB (B. subtilis)	1CSP_	67
CspA (E. coli)	3MEFA	69
PsaE	1PSF_	69
MerP	2HQI_	72
Tendamistat	2AIT_	74
Ubiquitin	1UBQ_	76
Protein L	2PTL_	78
Procarboxipeptidase A2	1PBA_	81
HPr	1HDN_	85
Im9	1IMQ	86

[illegible]	[illegible]	[illegible]
ACBP	2ABD_	86
PI3 SH3 domain	1PNJ_	86
FN_III (Domain 9)	1FNF_9	89
TNfn3	1TEN_	90
CTL9 (residues 58 to 149)	1DIV_	92
Monomeric λ-repressor	1LMB3	92
Twitchin	1WIT_	93
FN_III (Domain 10)	1FNF_10	93
U1A	1OIAA	95
CD2 (Domain 1)	1HNGA	97
Titin	1TIT_	98
mAcP	1APS_	98
Ribosomal protein S6	1RIS_	101
Horse Cytochrome c	1HRC_	104
Cytochrome b562	256BA	106
Rd-apocyt b562	1M6TA	106
Yeast Cytochrome c	1YCC_	107
FKBP12	1FKB_	107
Barnase	1A2PA	110
Villin 14T	2VIL_	126
Deoxymyoglobin	5MBN_	153
Cyclophilin A	1LOPA	164
VlsE	1L8WD	356

PDB, Protein Data Bank.

Table 2. N-element to C-element (N-C) contact probability in single domain protein chains

Single domain protein data set	N-C contact proteins	
	At least one contacting residue in each terminal element	**At least two contacting residues in each terminal element**
Less than 30% sequence similarity and 10 Å structure similarity (PDB-REPRDB):		
1543 (nss > 1)	772 (50%)	567 (37%)
1535 (nss > 1) (ß-hairpin as a terminal element)	835 (54%)	678 (44%)
1363 (nss > 3, seqres > 60 and seqsep > 30)	627 (46%)	455 (33%)
527 (nss > 12, seqres > 160 and seqsep > 30)	134 (25%)	82 (16%)
Less than 90% sequence similarity and 4 Å structure similarity (PDB-REPRDB):		
2308 (nss > 3, seqres > 60 and seqsep > 30)	1101 (48%)	821 (36%)
Less than 30% sequence similarity (PISCES):		
1374 (nss > 3, seqres > 60 and seqsep > 30)	636 (46%)	472 (34%)
Nonfragmented proteins (Less than 30% sequence similarity, PISCES):		
964 (nss > 3, seqres > 60 and seqsep > 30)	410 (43%)	298 (31%)

Unless otherwise indicated, a-helices and ß-strands were considered as single structural elements. Here nss represents number of secondary structural elements per protein, seqres represents number of residues per protein including unstructured residues, and seqsep represents residue sequence separation between terminal elements

Table 3. N-element to C-element (N-C) contact probability vs. random probability

Method of determining random probability	At least one contacting residue in each terminal element		At least two contacting residues in each terminal element	
	Random probability	**N-C contact/ random**	**Random probability**	**N-C contact/ random**
Terminal elements (N or C) contacting the middle element(s) in all single domain proteins	20%	2.3	11%	3.0
Same as above, but only in N-C contact single domain proteins *(conditional probability)*	23%	4.3	16%	6.4
Terminal elements (N or C) contacting the other penultimate element (C-1 or N+1 respectively) in N-C no-contact single domain proteins *(conditional probability)*	22%	4.7	17%	6.1
N-C contact probability in multi domain proteins	15%	3.1	7%	4.7
Terminal elements (N or C) contacting the middle element(s) in bigger single domain proteins (nss > 12 and seqres > 160)	13%	1.9	6%	2.5

Data set was selected at 30% sequence similarity and 10 Å structural similarity using PDB-REPRDB with nss > 3, seqres > 60 and seqsep > 30. All contacts between elements separated by less than two elements were excluded. Here, nss represents the number of secondary structural elements per protein, seqres represents the number of residues per protein including unstructured residues, and seqsep represents residue sequence separation between terminal elements.

Table 4. List of all N-element to C-element (N-C) no-contact proteins so far studied

Protein	PDB code	Length
N-terminal domain of HypF	1GXTA	91
Suc 1	1SCEC	113
Apo-pseudoazurin	1ADWA	123
Staphylococcal nuclease	1JOOA	149
P16 protein	2A5E_	156
GroEL apical domain	1AONA	157
DHFR	1RA9_	159
N-terminal domain of PGK	1PHP_	175
C-terminal domain of PGK	1PHP_	219
GFP	1B9CA	236
Trp synthase β2-subunit	1QOPB	396

All proteins show multistate folding kinetics. PDB, Protein Data Bank.

Fig. 7. Kekule's dream of a snake biting its own tail, which explained the structure of benzene and other ring systems, is mimicked in known protein structures. Essentially all two-state folding proteins and half of all single domain proteins have their N-terminal (blue) and C-terminal (red) secondary structural elements in contact (green). The figure shows six representative two-state proteins, in the order of increasing chain length: CI2 (2CI2I), ubiquitin (1UBQ_), mAcP (1APS_), cytochrome *c* (1HRC_), deoxymyoglobin (5MBN_), and VlsE (1L8WD) (generated by using MOLSCRIPT; ref. 1)

1. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24**, 945-949.