Protein Science (1995), *4*:1188–1202. Cambridge University Press. Printed in the USA.

Automatic recognition of hydrophobic clusters and their correlation with protein folding units

MICHEAL H. ZEHFUS

Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy and Department of Biochemistry, College of Biological Sciences, The Ohio State University, Columbus, Ohio 43210

(RECEIVED December 2, 1994; ACCEPTED March 15, 1995)

Abstract

A method is described to objectively identify hydrophobic clusters in proteins of known structure. Clusters are found by examining a protein for compact groupings of side chains. Compact clusters contain seven or more residues, have an average of 65% hydrophobic residues, and usually occur in protein interiors. Although smaller clusters contain only side-chain moieties, larger clusters enclose significant portions of the peptide backbone in regular secondary structure. These clusters agree well with hydrophobic regions assigned by more intuitive methods and many larger clusters correlate with protein domains. These results are in striking contrast with the clustering algorithm of J. Heringa and P. Argos (1991, *J Mol Biol 220*:151–171). That method finds that clusters located on a protein's surface are not especially hydrophobic and average only 3–4 residues in size.

Hydrophobic clusters can be correlated with experimental evidence on early folding intermediates. This correlation is optimized when clusters with less than nine hydrophobic residues are removed from the data set. This suggests that hydrophobic clusters are important in the folding process only if they have enough hydrophobic residues.

Keywords: compactness; folding intermediates; hydrophobicity; hydrophobic cores; hydrophobic clusters; protein folding

Since the seminal work of Kauzmann (1959) outlining the nature of the hydrophobic force, an implicit assumption in protein folding has been that proteins possess an interior core of hydrophobic residues and that the formation of this core is a major force in the folding process. Interest in hydrophobic regions has undergone a renaissance in recent years. In theoretical work, Dill has proposed that a protein's regular secondary structure is a natural outcome of burying residues in a hydrophobic core (Dill et al., 1993). In experimental work, both hydrogen exchange experiments and NOESY experiments have identified early folding intermediates that are frequently correlated with hydrophobic regions (Evans et al., 1991; Matouschek et al., 1992a, 1992b; Pan & Briggs, 1992; Gronenborn & Clore, 1994).

Although there is much interest in hydrophobic regions, there is no widely accepted method to identify or delineate them. Hydrophobic regions are usually identified by visual inspection, and unit definitions vary widely. Because each investigator identifies hydrophobic regions using different criteria, the proposed tie between hydrophobic regions and early folding intermediates must be viewed skeptically. Thus, it is important that an objective method be found to identify such regions.

This paper describes how hydrophobic regions can be objectively identified by locating compact clusters of side chains. Using compactness to locate such regions makes intuitive sense because hydrophobic clusters, like oil drops, will maximize interior interactions while minimizing their surface area.

Although compactness, not hydrophobicity, is the prime selection criterion, the discovered clusters are very hydrophobic, containing an average of 65% hydrophobic residues (the average protein in the study had 42% hydrophobic residues). Two structurally distinct hydrophobic regions, called clusters and cores, are observed. Clusters are smaller, with side-chain atoms located in the center of the unit and main-chain atoms located in the periphery. Cores are larger and enclose backbone atoms as elements of regular secondary structure in their interior.

If the hydrophobic clusters are filtered to remove units with less than nine hydrophobic residues, a good correlation is seen between the clusters and early folding intermediates. This supports the premise that hydrophobic regions are important in early folding events. It also suggests that there is a threshold be-

Reprint requests to: Micheal H. Zehfus, Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy and Department of Biochemistry, College of Biological Sciences, The Ohio State University, Columbus, Ohio 43210; e-mail: zehfus@compact.pharmacy.ohio-state.edu.

Abbreviations: BPTI, bovine pancreatic trypsin inhibitor; FOM, figure of merit; Z, coefficient of compactness; ζ, normalized coefficient of compactness.

low which hydrophobic clusters are not stable enough to seed the protein folding process. Further study of the correlation between hydrophobic clusters and early folding intermediates will help us better understand protein folding.

Results and discussion

Proteins derive a great deal of stability by clustering hydrophobic residues into regions where they are in contact with each other but have little contact with the solvent. A hydrophobic cluster should have a minimum surface area for its enclosed volume, because this would maximize the favorable hydrophobic-hydrophobic interactions while minimizing the unfavorable hydrophobic-solvent interactions.

Compactness, as defined by the Z parameter, finds units that have a minimum surface area for their enclosed volume (Zehfus & Rose, 1986). Thus, there should be a correlation between compactness and hydrophobic clusters. This correlation is surprisingly strong. The average compact cluster contains 65% hydrophobic residues, and only about 5% of the clusters contain less than 40% hydrophobic residues.

Cluster hydrophobicity is size dependent. Smaller compact clusters are about 70% hydrophobic, whereas larger ones are only 50% hydrophobic. This effect may be explained by the position of the clusters within the protein. Smaller clusters are generally buried in the protein interior, whereas the larger clusters have expanded to include some of the protein's surface and therefore are more hydrophilic.

The basic topography of compact clusters also varies with unit size. In small compact clusters, the side chains are clustered together in the center of the unit, with the main-chain atoms located around the periphery. When clusters have more than about 15 residues, main-chain atoms are no longer restricted to the periphery of the unit but become buried within the clusters in elements of regular secondary structure. Units with main-chain inclusions are easily identified either by visual inspection, or by the dramatic decreases in absolute Z values (increasing compactness) when main-chain atoms are included in compactness calculations.

Figure 1 shows the compact hydrophobic clusters and other parameters of interest for a dozen different proteins. In this figure, each side chain of a cluster is marked with a ■ or a ▒ at the residue's position in the protein, and each unit has been systematically named to portray its position in the overall hierarchy of compact units. This naming system is explained in the Methods section. Although this figure displays all compact units, those displayed in bold (with the ■ mark) have passed two additional filtering steps derived empirically to obtain the best fit between the compact clusters and experimental evidence for early folding units in these proteins (see the Methods).

The discontinuous nature of hydrophobic clusters initially makes this figure hard to interpret; however, certain patterns can be observed that make good structural sense. In β-strands, alternate residues point in opposite directions. The pattern ■ ■ ■ shows that residues 1, 3, and 5 are clustered together, most likely on one side of a β-strand. In some cases, like the β-sheet region of units I.B.2 and I.C of ribonuclease, one can observe two distinct clusters stabilizing opposite faces of a β-sheet.

Interactions with one side of a helix are seen in patterns like ■ ■■ ■, where residues 1, 4, 5, 8, and 9 all cluster on one face of the helix. As with β-strands, one can find cases where the same helix is stabilized by two different clusters (α-lactalbumin units I.A.2 and I.A.3).

Because hydrophobic clusters are usually discontinuous, the presence of continuous stretches of residues in larger units is puzzling. This puzzle is explained by the observation, made earlier, that larger units contain inclusions of regular secondary structure. Continuous stretches of residues in a cluster occur when a piece of regular secondary structure has been incorporated into the cluster's interior.

Because each protein is unique, the cluster structure and the correlation of clusters with early folding intermediates will now be discussed on a protein-by-protein basis.

Barnase

A large twisted β-sheet forms the basic framework for this 110-residue protein (Baudet & Janin, 1991). Core I consists of both sides of roughly 2/3 of this sheet. Cluster I.A, containing 76% hydrophobic residues, is on one side and I.B, with 67% hydrophobic residues, is on the other. Unit II, with only 47% hydrophobic residues, is on the same side of the sheet as I.B, but the twist in the sheet keeps units II and I.B separate from each other. The structures of barnase and many of its clusters and subclusters are presented in Figure 2 as an example of a typical protein and its clusters.

Fersht's group has developed a detailed model for the folding of barnase based on extensive analysis of folding mutations and proton exchange data (Fersht et al., 1992; Matouschek et al., 1992a, 1992b; Serrano et al., 1992a, 1992b, 1992c). This model identifies three hydrophobic regions called $core_1$, $core_2$, and $core_3$. These regions correspond quite well with cluster I.A, unit II and cluster I.B, respectively. $Core_1$ and $core_3$ are thought to form simultaneously, early in the folding process. This region would correspond nicely to the core I compact unit. $Core_2$ forms later, most likely because it is more hydrophilic in nature. Consistent with this explanation one finds that core I (the combination of $core_1$ and $core_3$) is both larger and contains more hydrophobic residues than unit II ($core_2$).

The goodness of fit between a hydrophobic cluster and a folding intermediate may be estimated in a number called the figure of merit. The FOM is determined by first calculating the percentage of residues in the hydrophobic cluster that are in or adjacent to residues in the folding intermediate, then determining the percentage of residues in the folding intermediate that are in or adjacent to residues of the hydrophobic cluster, and then averaging these two numbers together. The FOM of the combination of $core_1 + core_3$ correlated with hydrophobic core I is 66%. The FOM of unit II and $core_3$ is 61%.

Using Cα contact distances, Yanagawa has divided this protein into six folding modules: 1–24, 24–52, 52–73, 73–88, 88–98, and 98–110. These modules are thought to correlate with structural units encoded by exons (Yanagawa et al., 1993). Interestingly, none of these modules appear to contain a complete hydrophobic cluster, but a few hydrophobic clusters seem to be composed of combinations of modules (unit II, and clusters II.A and I.B).

Ubiquitin

The framework of this 76-residue protein is a β-barrel where two β-strands have been replaced with a single helix (Vijay-Kumar

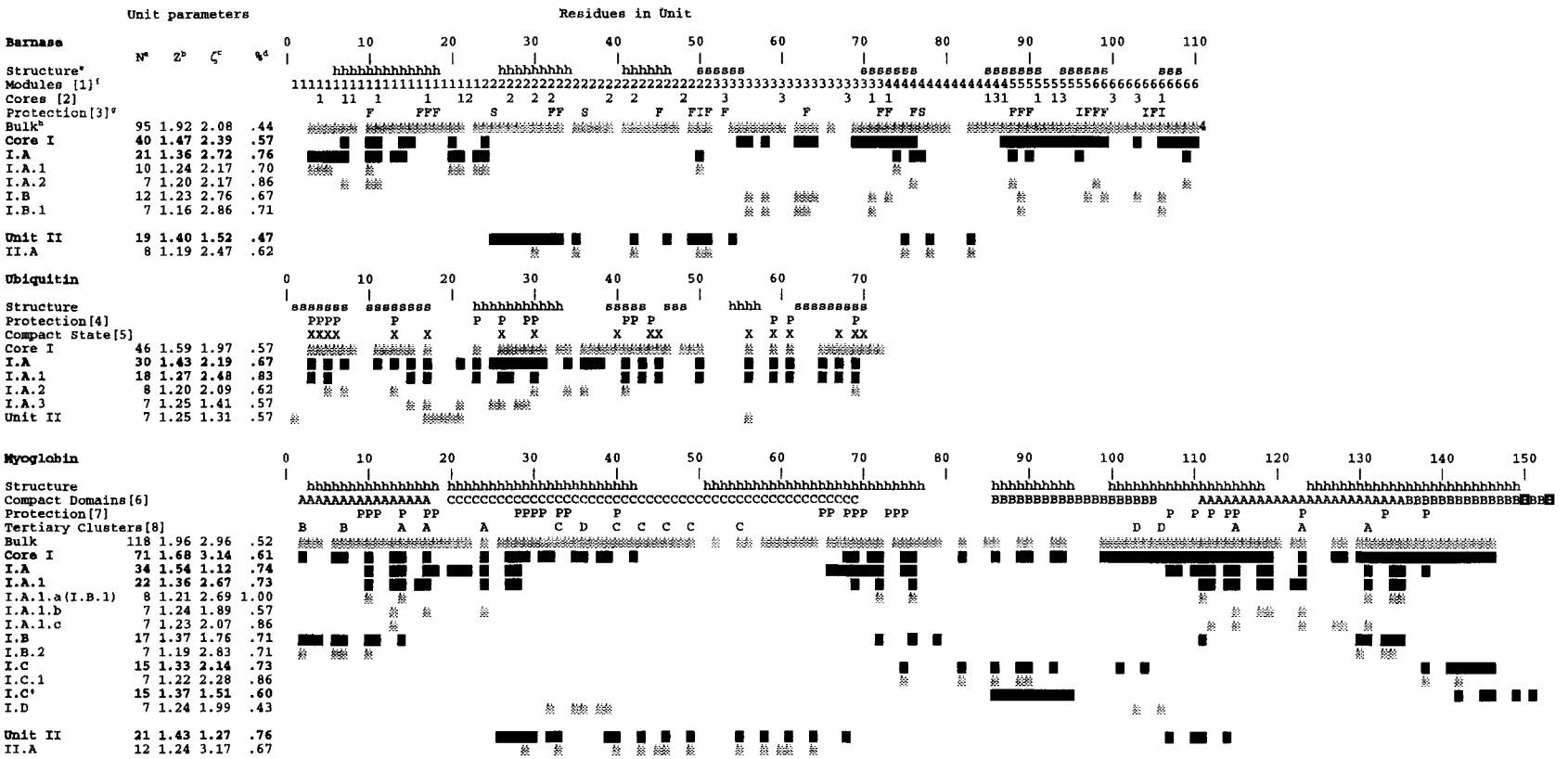

Fig. 1. Compact hydrophobic clusters from several proteins. [a]N, Number of residues in unit. [b]Z, *Z* of unit. [c]ζ, *Z* corrected for size effects. [d]%, % Hydrophobic residues. [e]h, Helix; s, sheet. [f]References: [1] Yanagawa et al. (193); [2] Serrano et al. (1992c); [3] Matouschek et al. (1992b); [4] Briggs and Roder (1992); [5] Pan and Briggs (1992); [6] Zehfus (1994); [7] Hughson et al. (1990); [8] Coco and Lecomte (1990); [9] Moore and Lecomte (1993); [10] Fisher and Taniuchi (1992); [11] Roder et al. (1988); [12] Jeng and Englander (1990); [13] Lu and Dahlquist (1992); [14] Miranker et al. (1991); [15] Buck et al. (1993); [16] Chyan et al. (1993); [17] Alexandrescu et al. (1993); [18] Udgaonkar and Baldwin (1990); [19] Mullins et al. (1993); [20] Varley et al. (1993); [21] Gronenborn and Clore (1994); [22] Staley and Kim (1990). [g]Rates for protection: F, fast, 12-50 s^{-1}; I, intermediate, 5-12 s^{-1}; S, slow, <5 s^{-1}. [h]Hierarchic designation of units described in the Methods. [i]P, Protected within 8 ms of folding; N, not protected in 8 ms. [j]W, weak; M, moderate; S, strong. [k]F, Fast, rate $\geq$ 25 s^{-1}; I, intermediate, rate < 25 s^{-1}. [l]E, Early protection, $t_{1/2}$ = 0.7-1.5 s; M, middle protection, $T_{1/2}$ = 15-25 s; L, long protection, $t_{1/2}$ > 25 s. (*Continued on the next three pages.*)

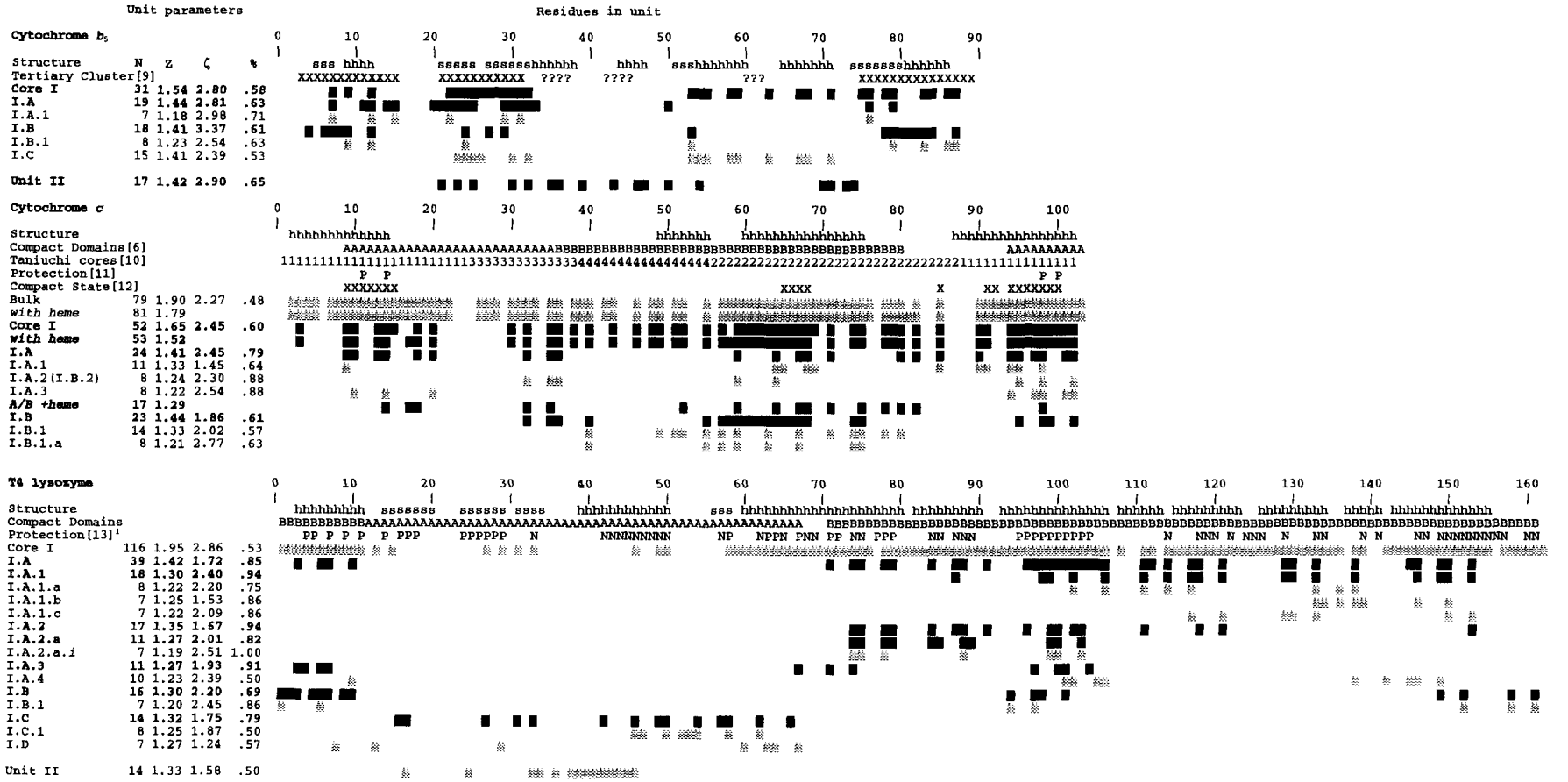

Fig. 1. *Continued.*

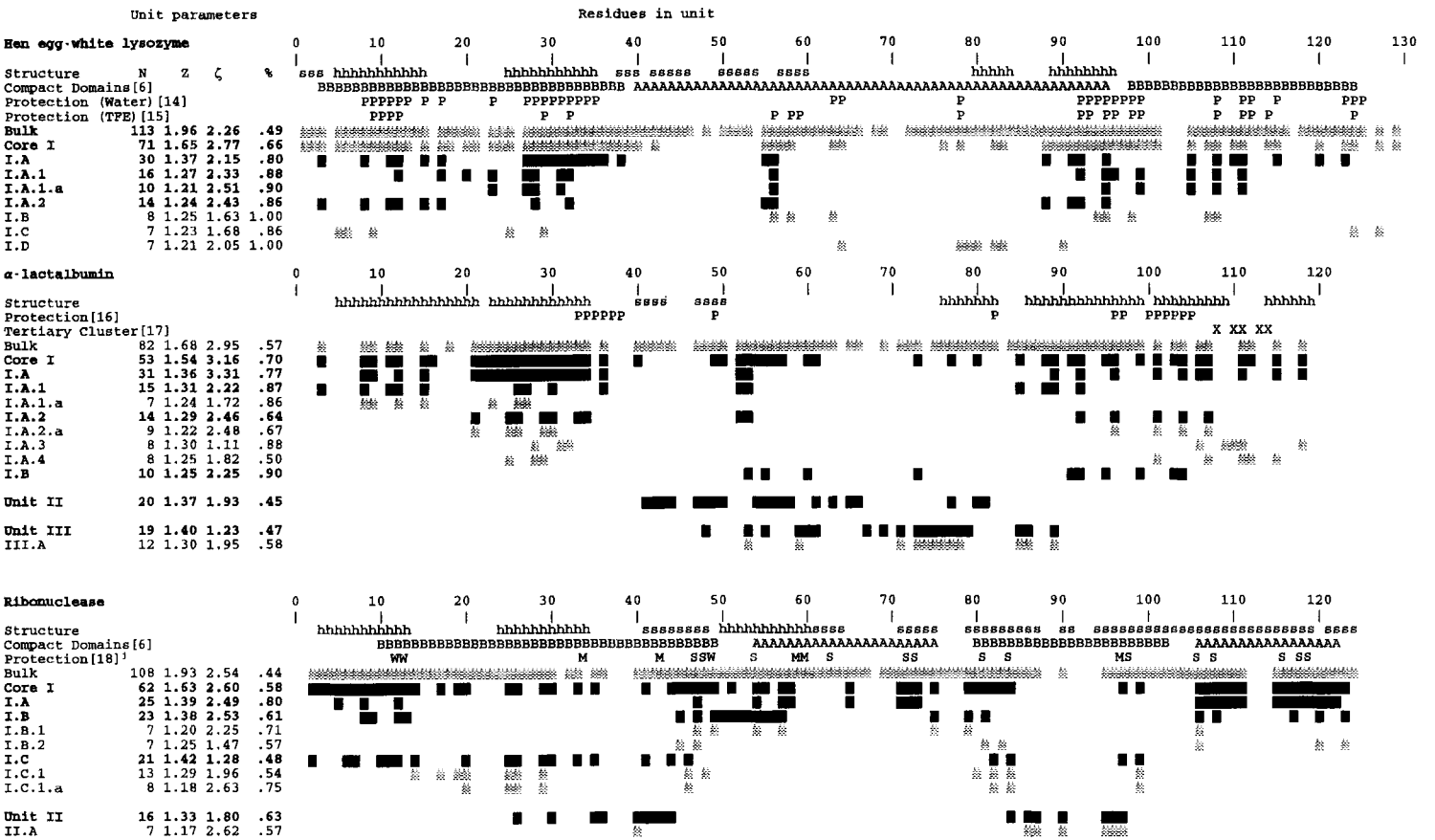

Fig. 1. *Continued.*

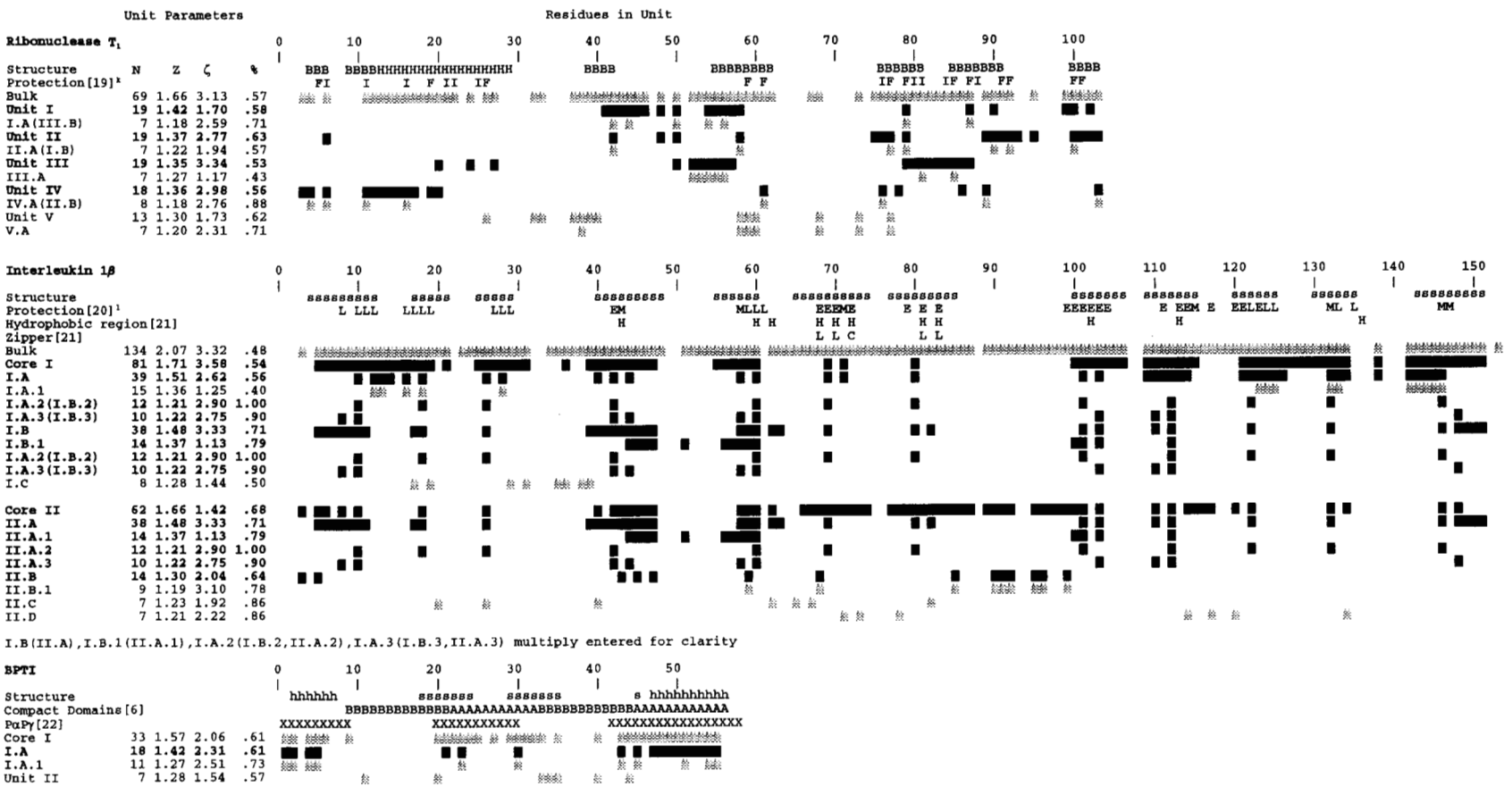

Fig. 1. *Continued.*

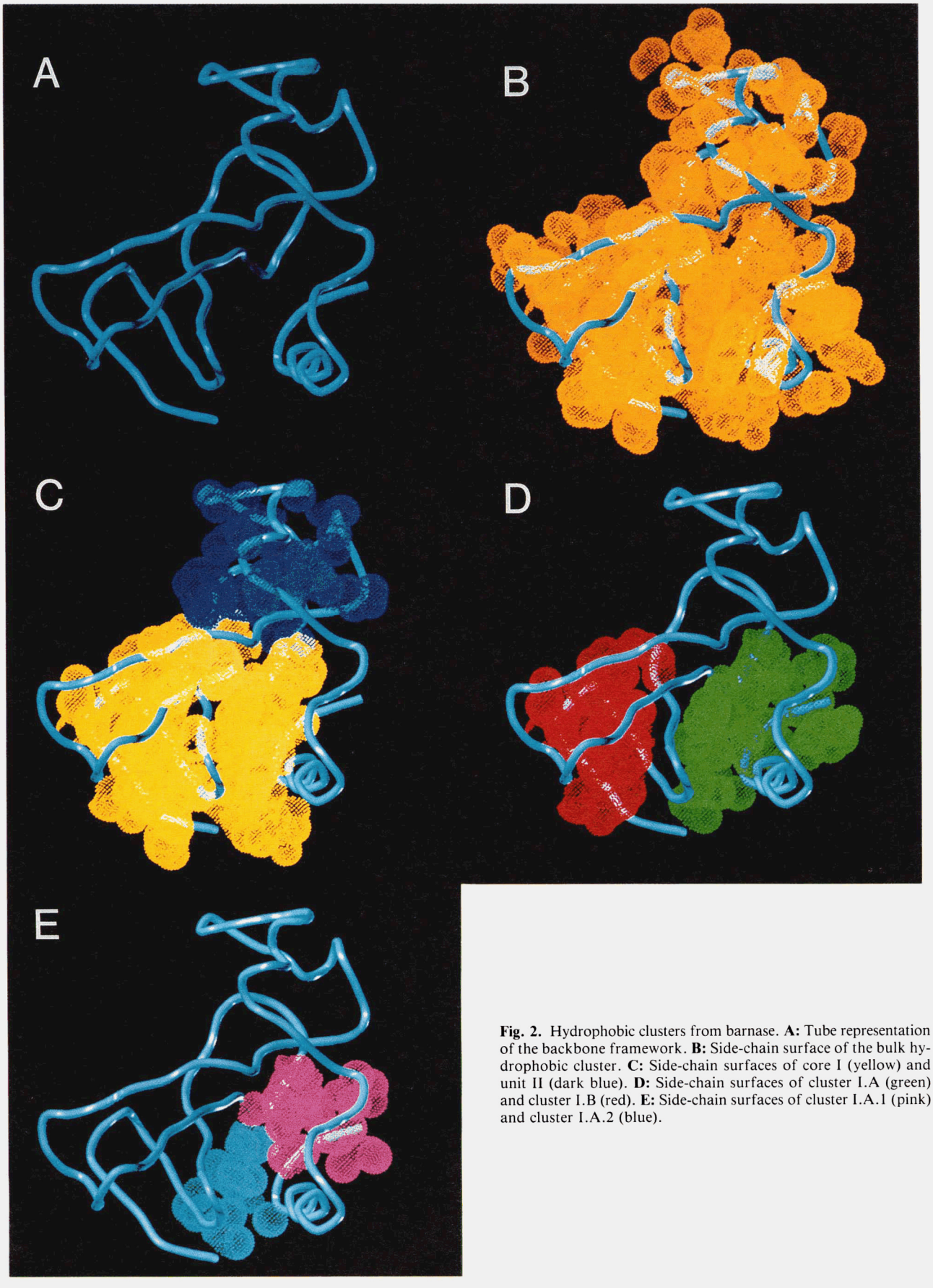

Fig. 2. Hydrophobic clusters from barnase. **A:** Tube representation of the backbone framework. **B:** Side-chain surface of the bulk hydrophobic cluster. **C:** Side-chain surfaces of core I (yellow) and unit II (dark blue). **D:** Side-chain surfaces of cluster I.A (green) and cluster I.B (red). **E:** Side-chain surfaces of cluster I.A.1 (pink) and cluster I.A.2 (blue).

et al., 1987). Hydrophobic cluster analysis finds a single large core I that corresponds well to this barrel. Cluster I.A is smaller, containing the core of the barrel and a few peripheral residues, and cluster I.A.1 represents the extremely hydrophobic center of the barrel. Clusters I.A.2 and I.A.3 overlap with I.A.1 but have little overlap with each other. They represent combinations of the I.A.1 central core with portions of a side or end of the barrel. Due to their large overlap with I.A.1, they are not viewed as separate structural entities, but as extensions of the central core. Unit II has little overlap with core I and represents a single strand of residues at one end of the barrel.

Using pulsed hydrogen exchange, Briggs and Roder found that this protein folds largely as a single unit, with most reporter protons being 80% protected from solvent exchange in the first 20 ms of folding (Briggs & Roder, 1992). This protein can form a compact non-native state at low pH in 60% methanol. This state has been studied by Pan and Briggs (1992), using pulsed hydrogen exchange experiments, and by Stockman et al. (1993), using heteronuclear techniques.

The exchange data, both in native and non-native states, correlate well with each other, identifying similar sets of protected nuclei. Stockman's data, which is restricted to amide proton NOE constraints, clearly shows the presence of the helix and two strands of β-sheet at the N-terminus of the protein but shows little distinct structure in the rest of the protein.

The hydrogen exchange data and the I.A.1 hydrophobic cluster are clearly closely related (FOM 81%). This cluster contains the hydrophobic residues from the middle of the barrel structure. The larger I.A structure is a superset of the I.A.1 cluster and contains additional residues from the exterior of the α-helix. Some of these additional residues are not protected from solvent exchange in the folding process, indicating that only the smaller cluster is important at this stage in folding.

Stockman's data identify a smaller intermediate in the A state containing only one helix and two strands of β-sheet. Apparently his methodology is more restrictive and fails to identify the more fluid regions of the nascent hydrophobic core.

Myoglobin

Myoglobin is a largely α-helical protein with a noncovalently bound heme group (Takano, 1984). To simplify calculations the heme was not included in the compact unit analysis. In this protein, core I is a very large entity that includes the bulk of the protein's interior. Core I buries most of the G and H helices and includes one face from all other helices except helix D.

Cluster I.A looks like a logical predecessor to core I. It has about 2/3 of the G helix completely buried and interactions are seen with faces of the A, B, and H helices. The I.A subclusters represent various helix-docking interactions within this unit.

This protein contains alternate structures, cluster I.C and I.C′. Alternate structures are places where the clustering algorithm identifies two units of approximately the same size and compactness, occupying roughly the same region of the protein. The choice between primary and alternate structure is made using hydrophobicity. The alternate structure is clearly much less hydrophobic than the primary unit.

Previously, compact domain analysis has been done on myoglobin (Zehfus, 1994). This analysis divided myoglobin into one continuous domain and two discontinuous domains. Good correlations are seen between unit II and the continuous domain and cluster I.C and one discontinuous domain. I.A shows moderate correlation with the second discontinuous domain.

When the heme is removed from myoglobin, its structure is largely disrupted, but some residual structure remains. The structure of apo-myoglobin has been probed using NMR both to quantitate proton exchange (Hughson et al., 1990) and to identify clusters of associating residues (Cocco & Lecomte, 1990). If compact hydrophobic clusters are important structural entities, then the same clusters found in the holoenzyme should also be observed in the experimental data from the apoenzyme.

Cocco and Lecomte (1990) identified four different clusters of hydrophobic residues in the apoenzyme. Two of these clusters, A and C in Figure 1, contain more than five residues, whereas the others are smaller. Cluster A corresponds well with the I.A.1 compact cluster (FOM 72%), whereas cluster C agrees only moderately with unit II (FOM 55%). No units are found that match the smaller clusters, but this is not surprising because they are too small to be detected with this methodology.

The NH protons protected from solvent exchange correlate quite well with cluster I.A (Hughson et al., 1990). Because this cluster is a larger version of the I.A.1 cluster that coincides with one of Lecomte's units, clearly compact hydrophobic clusters are important structural elements in both holomyoglobin and apomyoglobin.

Cytochrome b_5

Cytochrome b_5 is another protein that contains a noncovalently bound heme. Again, this heme is not included in the cluster analysis for simplicity. Unlike most other proteins in this analysis, cytochrome b_5 does not appear to have a large unit that includes most of the protein's residues. A compact unit of this size does exist, but it is not hydrophobic enough to be classed as a hydrophobic cluster.

This protein contains both helices and one large β-sheet (Mathews et al., 1972). Core I encompasses both sides of this β-sheet and represents the main core for more than half the molecule. Cluster I.A also includes residues from both sides of the β-sheet, although the bulk of this unit is on the side of the sheet distal from the heme. I.B also lies on the distal side of the sheet but represents a region where two helices dock with the β-sheet. I.C and unit II contact opposite faces of the heme. There is some overlap between these clusters along the edge of the heme closest to the β-sheet.

Like myoglobin, the heme of this protein may be removed to obtain a new, structurally different apoenzyme. The structure of the apoenzyme has been studied in detail by Moore and Lecomte (1993). The apocytochrome structural unit consists of a β-sheet and two short helices that are not involved in heme binding. This structural unit corresponds well with either the I.A or I.B clusters (FOMs 76% and 86%, respectively). Core I is a larger unit that contains both I.A and I.B, but it also includes residues from two additional helices that are not seen in the experimental data.

No correlation is seen with unit II, and this is puzzling because it is equivalent to both I.A or I.B in size, compactness, and hydrophobicity. Unit II, however, is in direct contact with the heme and probably requires some specific interactions for stabilization.

Cytochrome c

Cytochrome *c* has a covalently attached heme moiety. This makes the heme difficult to remove, and most experimental studies are done in its presence. Because the heme is present in the experimental work, a special analysis was done that included the heme, as well as the usual analysis where the heme is ignored. It can be seen that the presence of the heme significantly enhances the compactness of both the large bulk unit and core I. The heme can also fit into the smaller cluster that shares many elements of both the I.A and I.B apoclusters.

Core I is a large unit that encompasses most of the protein's interior. Cluster I.A sits on one end of the heme and makes contact between the N-terminal, 60s, 70s, and C-terminal helices and both faces of the heme. I.B represents just the docking of the 50s, 60s, and 70s helices with one face of the heme. When the heme is included in the analysis, a single 16-residue cluster is found that combines elements of both the I.A and I.B clusters with the heme to form a compact cluster. I.C shares several residues with I.B but is not in contact with the heme.

Clusters I.B and I.A.2 both correlate with compact domains found by Zehfus (1994) and seem to represent the hydrophobic cores of these domains. Taniuchi has identified four core domains in cytochrome by studying hybrid two-fragment complexes (Fisher & Taniuchi, 1992). Little correlation is seen with his domain regions.

The folding of this molecule with the bound heme has been studied using hydrogen-exchange labeling experiments (Roder et al., 1988). Ten protons in this protein's N- and C-terminal helices become 40% protected from solvent exchange in the first 30 ms of folding (Fig. 1 lists four that have been clearly identified). Additional folding steps then take place on 0.2-s and 10-s time scales. These two helices are also important in the structure of the non-native compact A state observed under low pH–high salt conditions. Here static hydrogen exchange experiments identify these two helices plus part of the 60s helix as a potential folding intermediate (Jeng et al., 1990).

The above helices are found in the I.A, I.B, and I.A/I.B–heme clusters of cytochrome *c*. Because the folding studies are done in the presence of the heme, the latter unit is the logical match. The FOM for this match is only 41%, indicating a poor fit between the units. This poor fit occurs because the I.A/I.B–heme cluster includes several residues not observed experimentally. Most of these additional residues are not in regions of regular secondary structure. Because protons in nonregular secondary structure exchange more quickly with the solvent, they are not detected easily in this kind of experiment. Thus, these additional residues may be part of the hydrophobic cluster, but this method cannot detect them.

T4 lysozyme

This is a good example of an $\alpha+\beta$ protein. One region in this protein is a large cluster of helices, and the second contains a single β-sheet with two auxiliary helices (Bell et al., 1991). This visual organization is not strictly observed in the hierarchy of hydrophobic clusters. Two smaller clusters are found in each domain, but no superclusters are found that correspond to the domains. Instead, core I includes the α domain and most of the β domain, excluding only a small flap of β into a separate unit II.

The α domain contains hydrophobic clusters I.A and I.B. I.A represents the true central core of this domain and almost completely buries the H5 helix (residues 93–106). I.B represents a bulge to one side of this core. These two clusters are in intimate contact, with 50% of the residues in I.B occurring in I.A. It is likely that the α domain has a single somewhat noncompact hydrophobic core that includes both I.A and I.B, and the cluster building algorithm divided this into two smaller but more compact subclusters. The β-sheet region of this protein is seen primarily in the I.C and unit II clusters. Unit I.D lies at the interface between the α and β region and may be included in either.

Early folding intermediates in T4 lysozyme have been characterized using pulsed hydrogen exchange (Lu & Dahlquist, 1992). These data show three regions protected from solvent exchange in the first 8 ms of folding. Many, but not all, of these residues occur in the small, highly hydrophobic I.A.3 cluster. Several other clusters (I.B, I.A.1, I.A.2) contain similar structural elements but combine them with other residues that are not protected at this time. Perhaps these structures fold more slowly and will only correlate with experimental data if the experimental conditions are varied so a later time point is observed.

The FOM of the I.A.3 cluster and the protected protons is 64%, indicating only a moderate correlation. The reason these structures are not better correlated is that a β-turn-β structure (residues 15–35) is protected from exchange but is not part of the I.A.3 cluster. These protected residues, in fact, do not occur in any hydrophobic cluster.

Visual inspection suggests that cluster I.C may stabilize this structure indirectly. The residues in question form a twisted β-turn-β structure. The I.C cluster is located at the base of the β-turn-β and keeps the strands locked together. This prevents the strands from separating and could greatly retard hydrogen exchange for protons within the strands. Thus, the collapse of a hydrophobic pocket may stabilize structures that lie outside the cluster itself.

Hen egg white lysozyme

This is another $\alpha+\beta$ protein (Diamond, 1974) where the large core I encompasses the bulk of the protein's interior, excluding a small flap of β structure, and includes both α and β domains. I.A is the true core of the α region and features a buried helix (helix B, residues 25–35). I.C is a bulge in the base of the α region away from the β-flap. There is only one residue in common between I.A and I.C, so it is not clear whether the core of this region is best represented by one large or two smaller independent clusters. The β region is stabilized by the two clusters I.B and I.D. These clusters are independent and have no residues in common.

Cluster I.A correlates reasonably well with a discontinuous compact domain found by Zehfus (1994). A second continuous domain from this analysis seems to use both I.B and I.D clusters as its hydrophobic core.

Miranker et al. (1991) have found 38 protons protected from hydrogen exchange in the first minute of this protein's folding. Buck et al. (1993) have studied a partially folded non-native compact form of this protein found in 50% trifluoroethanol at low pH. Evans et al. (1991) have studied thermally denatured lysozyme in aqueous solution and observed chemical shift effects consistent with the formation of hydrophobic clusters,

but only two specific interactions, 51–53 and 62–63, have been observed.

Figure 1 shows that both the region protected from solvent exchange early in folding and the region protected from solvent exchange in the compact non-native state correlate quite well with cluster I.A (FOMs 76% and 69%, respectively). The interactions observed in the thermally denatured state between residues 51 and 53 and residues 61 and 62 do not, however, play a prominent role in any native cluster.

α-Lactalbumin

Like the previous two proteins, α-lactalbumin contains two domains: one helical and the other a β-sheet and more random structure (Acharya et al., 1989). The helical region is well represented by core I.A. This cluster contains a single core helix (B, residues 23–34) that is almost completely buried by several other helices. The subsets of I.A represent different docking interactions between the core helix and the surrounding helices. Because several of these units overlap (I.A.3 and I.A.4; I.A.4 and I.A.2), it is likely that they do not represent independent folding units but are simply artificial subdivisions of the I.A unit.

Units II and III are spatially distinct from I.A and form the protein's second, irregular structural domain. Unit II encloses both sides of a β-turn-β between residues 40 and 50, whereas unit III is much more random with little secondary structure. Units II and III are about the same size and hydrophobicity and have about 40% overlap with each other. It is not clear whether they should be treated as small independent units or should be joined into one larger supercluster.

Cluster I.C lies at the interface between helical core I.A and the irregular unit II/unit III lobe. Cluster I.C is also consistent with the hydrophobic box used by Koga and Berliner (1985) to establish the structural similarity between this protein and lysozyme.

At low pH, α-lactalbumin goes into a compact non-native A-state. The A-state of guinea pig α-lactalbumin has been studied by hydrogen exchange (Chyan et al., 1993), and some side-chain clusters have been identified by NOESY interactions in the closely related bovine α-lactalbumin A-state (Alexandrescu et al., 1993). The correlation between hydrophobic domains and experimental data is not as good with this protein as it is for other proteins. The protection data correlate only moderately well (FOM 51%) with cluster I.A.2, and no compact clusters are found that correlate with the observed NOESY data.

One explanation for this is that the A-state of α-lactalbumin more closely resembles the denatured state than the native state. A major point of the Alexandrescu et al. (1993) paper is that no evidence is found for native-like hydrophobic clusters, and the large cluster observed is *not* consistent with the native structure.

Ribonuclease

This protein contains a single β-sheet, bent to form a V-like structure (Wlodawer et al., 1988). Again, core I is an oversized unit containing most of the protein except one small loop that forms the unit II cluster. The I.A and I.B clusters make one arm of the β-sheet. With 39% of the residues in I.B occurring in I.A, these two clusters overlap well and probably represent a single larger supercluster. The other arm of the β-sheet is stabilized by the I.C and unit II clusters. These two clusters also overlap (44% of unit II is part of cluster I.C) and should be joined into a second supercluster. There is only one residue in common between the I.A/I.B supercluster and the I.C/unit II supercluster, so these two units are independent of each other.

This protein has been divided into two binary discontinuous domains using compactness (Zehfus, 1994). One domain uses the I.C/unit II supercluster as its core, whereas a second domain uses parts of I.A/I.B as its core. Notice that the I.A/I.B supercluster contains residues from four distinct regions of ribonuclease. By definition, a binary domain contains residues from two regions, so it cannot be matched exactly to a cluster with four regions.

Udgaonkar and Baldwin (1990) followed the folding of ribonuclease using pulsed hydrogen exchange. Although most exchange experiments determine the time course of protection for individual protons, this experiment classes protons on the degree of protection (strong, medium, or weak) that exists after the first 0.4 s of folding. Most of these protected protons occur in both I.A and I.B clusters, confirming the hypothesis that these clusters should be joined in a single I.A/I.B supercluster. The FOM for the match between the I.A/I.B cluster and the protected residues is 66%. Few of the protected protons occur in the I.C/unit II supercluster. The preferred folding of cluster I.A/I.B over I.C/unit II is probably because the I.A/I.B cluster is larger (39 residues) and more hydrophobic (30 hydrophobics) than the unit II/I.C cluster (26 residues, 15 hydrophobics).

Ribonuclease T_1

The central skeleton of this molecule is a highly twisted β-sheet (Martinez-Oyanedel et al., 1991). There is no large single core in this protein, only five smaller clusters. Cluster I forms an inner core around which the β-sheet is twisted. Clusters III and IV are on the side of the sheet opposite cluster I and are located on different ends of an α-helix that stabilizes this side of the protein. Clusters II and V are located along the edges of the sheet.

Cluster I has one region that overlaps with cluster II and a second region that overlaps with cluster III. This suggests that the three clusters should be joined into a single I/II/III supercluster. Clusters IV and V, on the other hand, have little or no overlap with the other clusters and are largely independent. Cluster V, with only eight hydrophobic residues, is not hydrophobic enough to be a folding nucleation site.

The folding of ribonuclease T_1 is very complicated. Functionally, it is thought to be dependent on the *cis–trans* isomerization of two different proline amide bonds (Kiefhaber et al., 1990a, 1990b, 1990c). Experimentally, pulsed hydrogen exchange experiments detect two distinct folding phases thought to represent the folding of a native-like intermediate and a second non-native compact globular state (Mullins et al., 1993).

The residues associated with the faster folding form of ribonuclease (labeled F in Fig. 1) are thought to be associated with a native-like intermediate. Some of these residues can be found in each of the I, II, and III hydrophobic clusters, but no single cluster seems better correlated than the others. This is taken as evidence that these units should be grouped into a single supercluster rather than viewed as separate entities.

Interleukin

Interleukin's structure is that of a β-barrel where one end of the barrel has been closed off by trefoil arrangement of β-strands

(Veerapandian et al., 1992). Core I represents the interior of the β-barrel core plus several exterior residues. Core II is a bit smaller but again represents the central β-barrel core in combination with some exterior residues. The interiors of both these cores are almost identical, and the exteriors overlap but are not identical. The exterior residues shared by both cores are found in the I.B (or II.A) cluster family. The I.C cluster of core I continues from this surface in one direction, whereas the II.B and II.D clusters of core II continue in a different direction.

In this protein the cluster hierarchy is complicated because many higher order clusters share common subclusters. The subclusters that best represent the interior nucleus of the β-barrel are units I.A.2 and I.A.3. These units are used in three different superclusters (I.A, I.B, and II.A) and so have alternate designations I.B.2/I.B.3 and II.A.2/II.A.3. These two units share many residues (a 40% overlap) and are contiguous with each other. They almost certainly represent a larger, less compact cluster that has been artificially divided into two smaller pieces.

Pulsed hydrogen exchange techniques have been used to study the folding of interleukin (Varley et al., 1993). In contrast to the α-helical proteins, where NH protection may occur within a few milliseconds, protection here takes place in the 1–10-s time frame. The general pattern of protection correlates with the joint I.A/I.B cluster that represents the hydrophobic middle of the β-barrel. Although the general pattern of protection correlates with this combined cluster (FOM 85%), it is difficult to match the exact pattern of early, middle, and late protection to individual clusters or cores.

Gronenborn and Clore (1994) have proposed that folding of this protein is consistent with the hydrophobic zipper model of Dill et al. (1993). They further proposed that a set of four leucines and one cysteine formed the core of a "zipper" region. Although the evidence given here is consistent with hydrophobic regions being important in early folding events, none of the hydrophobic clusters found here correspond to the leucine zipper of Gronenborn and Clore.

Bovine pancreatic trypsin inhibitor

Core I of BPTI contains residues from the beginning, middle, and end of the protein's sequence and corresponds to the end of the molecule containing the α-helix. This cluster and its subclusters correspond quite closely to the PαPγ peptide used by Staley and Kim (1990) to model a two disulfide folding intermediate in BPTI. This peptide has native-like structure in solution, so the core I cluster family must play an important role in stabilizing this structure.

Both core I and unit II correspond to central regions of different discontinuous compact domains (Zehfus, 1994). Unit II is a good match because it is binary in nature. Core I, on the other hand, contains residues from three regions in BPTI and does not match exactly with the binary compact domain. The ternary hydrophobic cluster is probably the core of a true ternary compact domain, but the present domain identification scheme is not sophisticated enough to identify such units.

General trends

The hydrophobic clusters found here are intuitively pleasing. Smaller units come from the interior of the protein, are largely hydrophobic, and exclude regions of hydrogen bonded secondary structure. Larger units can be assembled by joining or expanding smaller units, are less hydrophobic, and enclose elements of regular secondary structure.

Although the hydrophobic clusters are intuitively pleasing, their structural hierarchies do not always match our intuitive picture of domains. Occasionally the clustering algorithm will expand a cluster beyond its domain boundary to include a significant portion of a neighboring domain. This is a minor problem because these situations are easily identified by visual inspection of the proposed units.

This method will also occasionally subdivide hydrophobic clusters into two or more smaller but more compact subclusters. These cases are identified by examining the overlap between units; whenever two clusters are roughly the same size and have more than 30% overlapping residues, they should be examined carefully to see if they should be joined into a larger supercluster. Several cases were shown where two, or even three, overlapping clusters were proposed to be a single larger supercluster. In these cases, the experimental evidence did not favor any single cluster but indicated that the clusters form simultaneously as a single unit.

Heringa and Argos (1991) have proposed a different method for finding densely packed clusters of side chains. Their clusters are strikingly different from the clusters found here. In the Heringa method, one exhaustively searches all side-chain combinations for pairs of side chains with maximal contact. After normalization for side-chain size, a threshold is set, and a set of "dense neighbor" pairs is selected. These pairs are then used to find dense clusters of three residues, and three-residue clusters are used to find four-residue clusters, etc. At each step of cluster growth, the added residue is checked for minimal contact with other residues in the cluster and for contact with residues outside the cluster. If the added residue has too much contact with residues outside the cluster, it is not added. The latter rule is used to make the clusters relatively isolated.

The clusters discovered by the Heringa method are usually located on a protein's surface and contain an average of three or four residues. As might be expected for surface residues, the clusters are not especially hydrophobic and frequently include charged residues. This contrasts sharply with the clusters found here, which are much larger, extremely hydrophobic, and usually located in a protein's interior.

Because both methods attempt to find densely packed clusters, and both methods use some form of surface area measurement to find these clusters, the difference in results is quite surprising. There are two likely explanations for this dissimilarity. First, for technical reasons, the compactness method cannot properly evaluate the compactness of small units (see the Methods) and only reliably locates units containing seven or more residues. Thus, this method cannot analyze the very small clusters observed by Heringa and may be examining an entirely different type of cluster.

The second explanation is that the Heringa method has an internal bias for surface residues. That method deliberately selects units that have little overlap with other residues. Such units are more likely to occur at a protein's surface, where one side of the unit is solvent exposed, than in the protein interior, where the unit's entire surface will be in contact with other residues. It remains for future investigations to decide which kind of cluster is structurally more interesting.

Recently Swindells (1995) described a method for finding hydrophobic clusters by identifying buried residues in regular secondary structures that have a large number of contacts between hydrophobic atoms. This approach is used to find a single core in a protein and does not yield the hierarchical cluster of families observed here. The core definition is also quite restrictive, so the largest observed core has only about 25 residues.

Three of the proteins analyzed by Swindells are also analyzed here. With a FOM of 88%, Swindells's core for interleukin 1β is an excellent match with the I.A.1 cluster found here. The core he chooses for myoglobin is also similar to the I.A.1 cluster found here, although its FOM is lower (67%). The biggest difference is in pancreatic trypsin inhibitor. Because the Swindells method uses a threshold accessibility value, only four residues in BPTI can be used to form a core. Because these residues are not in contact with each other, he finds no acceptable core for this protein. This contrasts with the four clusters found here, one of which is thought to be a significant folding intermediate.

Early folding intermediates are thought to correlate with hydrophobic regions in proteins. To see if this is true, hydrophobic clusters were correlated against experimentally identified folding intermediates. In nine cases the correlation is very good, and in two cases it is only moderate. Interestingly, the best correlation between hydrophobic clusters and folding intermediates occurs when clusters with less than nine hydrophobic residues are removed from the data set. This suggests that there is a minimum size for a hydrophobic pocket and that clusters below this threshold are not stable enough to initiate folding.

The total number of hydrophobic residues also appears to be important in determining relative stability between units. Several cases are seen where two or more possible clusters exist in a protein. In these cases the cluster with the most hydrophobic residues usually has the best correlation with the experimental data. This makes intuitive sense because the larger clusters would hide more hydrophobic surface area and be more energetically favored.

Correlating hydrophobic clusters to specific steps in folding is more difficult. In proteins where distinct ranges of protection or distinct phases of protection can be observed, it is usually not possible to tie individual clusters to distinct folding entities. This often occurs because a single structural entity, like a β-sheet, may be stabilized by more than one hydrophobic cluster. In these cases, it is not clear whether a single cluster is necessary for the structure to form, or if both clusters must coalesce simultaneously.

The cases where hydrophobic clusters and folding intermediates do not correlate may be explained in several ways. First, many experiments used to monitor folding intermediates are not well suited for observing tertiary folding interactions. For instance, hydrogen exchange labeling experiments usually only monitor residues in regions of regular secondary structure. If compact clusters contain residues outside helices or sheets, they will not be observed using this method. Further, the reactions involved in the exchange process are complex, and other factors, such as solvent accessibility, are also important in determining the proton exchange rates.

Making correlations between hydrophobic clusters and folding intermediates are particularly difficult in β-sheet regions. This is due to the discontinuous nature of the β-sheet. The hydrophobic cluster that stabilizes a β-sheet may contain only one or two residues from each β-strand. In this case, the entire sheet is stabilized by a few key cluster residues, so the experimentally observed structure may be much larger than the cluster responsible for the structure. To make matters even more confusing, a single sheet may be stabilized by two different clusters, one on each side of the sheet. Thus, a simple one-to-one correspondence between protected residues and clusters will almost never occur in β regions.

Another factor that may preclude a correlation between hydrophobic clusters and proposed folding intermediates is that the structure being analyzed may not be a folding intermediate. In a few proteins the compact non-native A-state was used as a folding intermediate model. In cytochrome *c* and ubiquitin, this is reasonable because the proton exchange data of the A-state resemble the proton exchange data in the native state. In α-lactalbumin, however, there is no such direct evidence tying the "molten globule" state directly to the folding process, and indeed, there is good evidence that this particular "molten globule" has some very non-native features (Alexandrescu et al., 1993).

Considering all the factors that can cloud the correlation between hydrophobic clusters and folding intermediates, it is actually quite encouraging that the correlation is as good as it is. It appears that compact hydrophobic clusters are important in early folding intermediates, and that cluster size can be directly tied to a unit's existence and stability.

Methods

Compact hydrophobic clusters were found for the proteins α-lactalbumin 1ALC (Acharya et al., 1989), barnase 1RNB (Baudet & Janin, 1991), ubiquitin 1UBQ (Vijay-Kumar et al., 1987), myoglobin 4MBN (Takano, 1984), cytochrome b_5 3B5C (Mathews et al., 1972), cytochrome *c* 3CYT (Takano & Dickerson, 1980), interleukin 4I1B (Veerapandian et al., 1992), T4 lysozyme 4LZM (Bell et al., 1991), hen egg-white lysozyme 6LYZ (Diamond, 1974), ribonuclease 7RSA (Wlodawer et al., 1988), BPTI 6PTI (Wlodawer et al., 1987), and ribonuclease T_1 9RNT (Martinez-Oyanedel et al., 1991), using coordinates deposited in the Brookhaven Protein Data Bank (Bernstein et al., 1977). In all cases, heteroatoms and acetylations were removed from the coordinate sets. If a side-chain position was multiply defined, the single most populated side-chain position was used.

An exception to the above was made for cytochrome *c*. This protein contains a single covalently bound heme group, and refolding and protection studies done on this protein include the heme moiety. To see if the presence of the heme changes the analysis, the analysis was performed both with and without the heme group.

Compactness, measured by dividing an object's solvent-accessible surface area by its minimum possible surface area, has been used to locate continuous and binary discontinuous protein domains (Zehfus, 1987, 1993, 1994). In the search for domains, main-chain and side-chain moieties of an amino acid are considered as one inseparable unit, and domains are assembled continuously, by adding new amino acids at either end of a growing, continuous peptide segment.

In the search for compact clusters, a much different approach is taken. First, because the peptide backbone is inherently hydrophilic and should not appear in a hydrophobic region until its polar moieties are hydrogen bonded, *side chains only* are used in building clusters. Backbone moieties may be added to these

clusters in later steps of the analysis for clarification of backbone interactions, but they are not used directly in the cluster-building algorithm.

The second difference between this method and the domain-finding algorithm is that side chains are added *discontinuously*. This is necessary because clusters are inherently discontinuous; a cluster on one side of a β-sheet will only contain every other residue within a single β-strand, and the strands of a β-sheet frequently are not contiguous with each other. Clusters formed by the docking of helices are also inherently discontinuous, although the pattern of the discontinuity will be different.

Because we wish to discover hydrophobic clusters, it is natural to think that we should limit our search to only hydrophobic side chains. It is impossible, however, to class many side chains as purely hydrophobic or hydrophilic because they contain a mixture of properties. Consequently, all side chains are used in the discovery algorithm. It will be seen that even when hydrophilic side chains are included, the discovered units are very hydrophobic. A few hydrophilic clusters are found, but they are easily identified and removed.

The gist of the method is that each side chain in a protein is used to grow a family of potential clusters called a "clump." Clumps are grown in a stepwise manner, by finding the single side chain that optimizes the compactness of the growing clump. Once clumps have been identified for each side chain in a protein, they are compared with each other to find the sets of side chains that are most compact. The most compact side-chain sets are then identified as compact clusters or cores.

Details of the cluster-finding algorithm

A clump is started from a given side chain by pairing it with all other side chains in the protein and evaluating each pair's compactness. When the most compact pair is found, the identity of the added side chain and the compactness of the pair are recorded, and the first step in clump growth is complete. Next, this pair of side chains is then tried in combination with all remaining side chains to find the most compact side-chain triplet. This process is continued, and the clump grows, until all of the protein's side chains are contained in the clump.

During this process, compactness is evaluated with the Z parameter (Zehfus & Rose, 1986), using a look-up table-based algorithm for area and volume calculations (Zehfus, 1993), specially modified to use only side-chain atoms. The number of calculations done can be greatly reduced if clumps are inspected after each step of growth to remove clumps with identical sets of side chains.

The next step in the procedure is to examine the clumps for sets of side chains that are exceptionally compact. Comparing the compactness of units of different sizes is difficult, however, because Z values for these units are size dependent (data not shown). A size-independent parameter called ζ was developed to remove this complication. ζ is simply the number of standard deviations below the mean of a given Z (compactness) value.

At each step of clump growth, a record is kept of all calculated compactness values, so the mean and standard deviation of the compactness distribution are known and can be used to calculate ζ. For example, core I from ubiquitin has an absolute Z value of 1.59 and a ζ of 1.97. The ζ value tells us that this unit is almost two standard deviations below the mean Z value for units of this size within this protein. By comparing distance from the mean Z value, rather than absolute Z value, ζ becomes a size-independent parameter for comparing compactness values.

Using the ζ parameter, each clump is now searched for a few side-chain sets that are very compact. To do this, each member in the clump is compared to other, similar-sized members of the clump. If a given unit has N residues, and another unit containing between N/2 to 3N/2 residues is found with a higher ζ value, the unit with the lower ζ is rejected.

Many of the remaining units are closely related, describing nearly identical pieces of structure. To find the best unit to represent each structure, each cluster is next compared to other similar-sized clusters to see if they are closely related. If the potential cluster has N residues, then all units between N/2 and 3N/2 residues are searched for sequence similarities. If more than 50% of a smaller unit's residues occur within a larger unit, the clusters are declared similar, and only the cluster with the highest ζ values is retained.

Although hydrophobicity is not used as a constraint in finding these compact clusters, the final units are largely hydrophobic. When alanine, valine, leucine, isoleucine, proline, phenylalanine, tryptophan, methionine, cystine, and tyrosine are considered hydrophobic, the average compact cluster contains 65% hydrophobic residues, whereas the average protein used in this study contains only 42% hydrophobic residues. The procedure does, however, find some clusters that are not hydrophobic. These hydrophilic clusters are removed by eliminating all clusters with less than 40% hydrophobic residues. Nearly 95% of all compact clusters pass this test.

All discovered hydrophobic clusters are listed in Figure 1. When all compact hydrophobic clusters are compared with the experimental data for early folding intermediates, it becomes apparent that neither the largest nor the smallest clusters correlate with folding intermediates. This was investigated further to try to understand this phenomenon, and to design data filters that would remove noncorrelating units from the data set.

The largest clusters are much bigger than the units delineated by the experimental data and probably represent structures formed late in the folding process. These large units are removed from the data sets by eliminating all clusters containing more than 50% of a protein's amino acids.

The most plausible reason that smaller units do not correlate with observed structure is that nascent units must exceed some threshold energy value before they are stable enough to be observed. To try to understand this better, the set of hydrophobic clusters was filtered using size, percent hydrophobicity, compactness, or total number of hydrophobic residues, to see which parameter would give the best fit with experimental data. The best correlation was found with total number of hydrophobic residues. Units containing nine or more hydrophobic residues correlate well with experimental data, whereas those with eight or fewer hydrophobics have little or no correlation. The clusters in Figure 1 that pass these additional filtering steps are printed in bold type.

To help organize the clusters, Figure 1 lists units in natural hierarchies. To establish these hierarchies, units were first sorted by size, and then the amount of each smaller unit that exists within a larger unit was calculated. If more than 50% of a smaller unit is included in a larger unit, the smaller is considered a subset of the larger. The hierarchical relationships between clusters are summarized in the notation used in column 1 of Figure 1. In this notation each cluster is given a designa-

tion of numbers and letters, such as I.B.2.a. Much like the letters in an outline, the first numeral designates the largest independent structure this unit appears in, and each additional number or letter indicates a smaller and smaller subset of the larger unit. Occasionally two slightly different clusters occupy roughly the same region of a protein. In this case, ′ is used to designate the alternate definition, as in I.A′. Other times, the same substructure can be used in different superstructures. In this case, the designation using the alternate superstructure nomenclature is added in parentheses. This method is not foolproof, and visual inspection of clusters is sometimes necessary to resolve ambiguities.

The largest cluster in a protein usually contains more than half of the protein's residues and has all the other clusters as subclusters. This large supercluster is termed "bulk" because it contains the bulk of the protein's interior residues. The next largest cluster is termed either core I or unit I, depending on its size. Clusters with more than 25 residues are clearly major structural entities and are called cores. Clusters with less than 25 residues are called units because their structural import is less certain. The largest cluster that is not a subset of core I or unit I is termed core or unit II.

Acknowledgments

I thank George Rose for thoughtful comments and the National Institutes of Health (grant GM46664) for supporting this work.

References

Acharya KR, Stuart DI, Walker NP, Lewis M, Phillips DC. 1989. Refined structure of baboon alpha-lactalbumin at 1.7 Å resolution. Comparison with C-type lysozyme. *J Mol Biol 208*:99–127.

Alexandrescu AT, Evans PA, Pitkeathly M, Baum J, Dobson CM. 1993. Structure and dynamics of the acid-denatured molten globule state of alpha-lactalbumin: A two-dimensional NMR study. *Biochemistry 32*: 1707–1718.

Baudet S, Janin J. 1991. Crystal structure of a barnase-d(GpC) complex at 1.9 Å resolution. *J Mol Biol 219*:123–132.

Bell JA, Wilson KP, Zhang XJ, Faber HR, Nicholson H, Matthews BW. 1991. Comparison of the crystal structure of bacteriophage T4 lysozyme at low, medium, and high ionic strengths. *Proteins Struct Funct Genet 10*:10–21.

Bernstein FC, Koetzle TG, Williams GJB, Meyer EF Jr, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structure. *J Mol Biol 122*:535–542.

Briggs MS, Roder H. 1992. Early hydrogen-bonding events in the folding reaction of ubiquitin. *Proc Natl Acad Sci USA 89*:2017–2021.

Buck M, Radford SE, Dobson CM. 1993. A partially folded state of hen egg white lysozyme in trifluoroethanol: Structural characterization and implications for protein folding. *Biochemistry 32*:669–678.

Chyan CL, Wormald C, Dobson CM, Evans PA, Baum J. 1993. Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: A hydrogen exchange study. *Biochemistry 32*:5651–5691.

Cocco MJ, Lecomte JT. 1990. Characterization of hydrophobic cores in apomyoglobin: A proton NMR spectroscopy study. *Biochemistry 29*:11067–11072.

Diamond R. 1974. Real-space refinement of the structure of hen egg-white lysozyme. *J Mol Biol 82*:371–391.

Dill KA, Fiebig KM, Chan HS. 1993. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci USA 90*:1942–1946.

Evans PA, Topping KD, Woolfson DN, Dobson CM. 1991. Hydrophobic clustering in nonnative states of a protein: Interpretation of chemical shifts in NMR spectra of denatured states of lysozyme. *Proteins Struct Funct Genet 9*:248–266.

Fersht AR, Matouschek A, Serrano L. 1992. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol 224*:771–782.

Fisher A, Taniuchi H. 1992. A study of core domains, and the core domain-domain interaction of cytochrome *c* fragment complex. *Arch Biochem Biophys 296*:1–16.

Gronenborn AM, Clore GM. 1994. Experimental support for the "hydrophobic zipper" hypothesis. *Science 263*:536.

Heringa J, Argos P. 1991. Side-chain clusters in protein structures and their role in protein folding. *J Mol Biol 220*:151–171.

Hughson FM, Wright PE, Baldwin RL. 1990. Structural characterization of a partly folded apomyoglobin intermediate. *Science 249*:1544–1548.

Jeng MF, Englander SW, Elove GA, Wand AJ, Roder H. 1990. Structural description of acid-denatured cytochrome *c* by hydrogen exchange and 2D NMR. *Biochemistry 29*:10433–10437.

Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. *Adv Protein Chem 14*:1–64.

Kiefhaber T, Grunert HP, Hahn U, Schmid FX. 1990a. Replacement of a cis proline simplifies the mechanism of ribonuclease T_1 folding. *Biochemistry 29*:6475–6480.

Kiefhaber T, Quaas R, Hahn U, Schmid FX. 1990b. Folding of ribonuclease T_1. 1. Existence of multiple unfolded states created by proline isomerization. *Biochemistry 29*:3053–3061.

Kiefhaber T, Quaas R, Hahn U, Schmid FX. 1990c. Folding of ribonuclease T_1. 2. Kinetic models for the folding and unfolding reactions. *Biochemistry 29*:3061–3070.

Koga K, Berliner LJ. 1985. Structural elucidation of a hydrophobic box in bovine alpha-lactalbumin by NMR: Nuclear Overhauser effects. *Biochemistry 24*:7257–7262.

Lu J, Dahlquist FW. 1992. Detection and characterization of an early folding intermediate of T4 lysozyme using pulsed hydrogen exchange and two-dimensional NMR. *Biochemistry 31*:4749–4756.

Martinez-Oyanedel J, Choe HW, Heinemann U, Saenger W. 1991. Ribonuclease T_1 with free recognition and catalytic site: Crystal structure analysis at 1.5 Å resolution. *J Mol Biol 222*:335–352.

Mathews FS, Argos P, Levine M. 1972. The structure of cytochrome b_5 at 2.0 Ångstrom resolution. *Cold Spring Harbor Symp Quant Biol 36*: 387–395.

Matouschek A, Serrano L, Fersht AR. 1992a. The folding of an enzyme. IV. Structure of an intermediate in the refolding of barnase analysed by a protein engineering procedure. *J Mol Biol 224*:819–835.

Matouschek A, Serrano L, Meiering EM, Bycroft M, Fersht AR. 1992b. The folding of an enzyme. V. H/2H exchange-nuclear magnetic resonance studies on the folding pathway of barnase: Complementarity to and agreement with protein engineering studies. *J Mol Biol 224*:837–845.

Miranker A, Radford SE, Karplus M, Dobson CM. 1991. Demonstration by NMR of folding domains in lysozyme. *Nature 349*:633–636.

Moore CD, Lecomte JT. 1993. Characterization of an independent structural unit in apocytochrome b_5. *Biochemistry 32*:199–207.

Mullins LS, Pace CN, Raushel FM. 1993. Investigation of ribonuclease T_1 folding intermediates by hydrogen-deuterium amide exchange-two-dimensional NMR spectroscopy. *Biochemistry 32*:6152–6156.

Pan Y, Briggs MS. 1992. Hydrogen exchange in native and alcohol forms of ubiquitin. *Biochemistry 31*:11405–11412.

Roder H, Elöve GA, Englander SW. 1988. Structural characterization of folding intermediates in cytochrome *c* by H-exchange labelling and proton NMR. *Nature 335*:700–704.

Serrano L, Kellis JT Jr, Cann P, Matouschek A, Fersht AR. 1992a. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J Mol Biol 224*:783–804.

Serrano L, Matouschek A, Fersht AR. 1992b. The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J Mol Biol 224*:805–818.

Serrano L, Matouschek A, Fersht AR. 1992c. The folding of an enzyme. VI. The folding pathway of barnase: Comparison with theoretical models. *J Mol Biol 224*:847–859.

Staley JP, Kim PS. 1990. Role of a subdomain in the folding of bovine pancreatic trypsin inhibitor. *Nature 344*:685–688.

Stockman BJ, Euvrard A, Scahill TA. 1993. Heteronuclear three-dimensional NMR spectroscopy of a partially denatured protein: The A-state of human ubiquitin. *J Biomol NMR 3*:285–296.

Swindells MB. 1995. A procedure for the automatic determination of hydrophobic cores in protein structure. *Protein Sci 4*:93–103.

Takano T. 1984. Refinement of myoglobin and cytochrome *c*. In: Hall SR, Ashida T, eds. *Methods and applications in crystallographic computing*. Oxford, UK: Oxford University Press. pp 262–272.

Takano T, Dickerson RE. 1980. Redox conformation changes in refined tuna cytochrome *c*. *Proc Natl Acad Sci USA 77*:6371–6373.

Udgaonkar JB, Baldwin RL. 1990. Early folding intermediate of ribonuclease A. *Proc Natl Acad Sci USA 87*:8197–8201.

Varley P, Gronenborn AM, Christensen H, Wingfield PT, Pain RH, Clore GM. 1993. Kinetics of folding of the all-beta sheet protein interleukin-1 beta. *Science 260*:1110–1113.

Veerapandian B, Gilliland GL, Raag R, Svensson AL, Masui Y, Hirai Y, Poulos TL. 1992. Functional implications of interleukin-1 beta based on the three-dimensional structure. *Proteins Struct Funct Genet 12*:10–23.

Vijay-Kumar S, Bugg CE, Cook WJ. 1987. Structure of ubiquitin refined at 1.8 Å resolution. *J Mol Biol 194*:531–544.

Wlodawer A, Nachman J, Gilliland GL, Gallagher W, Woodward C. 1987. Structure of form III crystals of bovine pancreatic trypsin inhibitor. *J Mol Biol 198*:469–480.

Wlodawer A, Svensson LA, Sjolin L, Gilliland GL. 1988. Structure of phosphate-free ribonuclease A refined at 1.26 Å. *Biochemistry 27*:2705–2717.

Yanagawa H, Yoshida K, Torigoe C, Park JS, Sato K, Shirai T, Go M. 1993. Protein anatomy: Functional roles of barnase module. *J Biol Chem 268*:5861–5865.

Zehfus MH. 1987. Continuous compact protein domains. *Proteins Struct Funct Genet 2*:90–110.

Zehfus MH. 1993. Improved calculations of compactness and a re-evaluation of continuous compact units. *Proteins Struct Funct Genet 16*:293–300.

Zehfus MH. 1994. Binary discontinuous compact domains. *Protein Eng 7*:335–340.

Zehfus MH, Rose GD. 1986. Compact units in proteins. *Biochemistry 25*: 5759–5765.