Protein Science (1997), 6:1210–1219. Cambridge University Press. Printed in the USA.

Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains

MICHEAL H. ZEHFUS

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

(RECEIVED August 28, 1996; ACCEPTED March 14, 1997)

Abstract

Compactness has been used to locate discontinuous structural units containing one or more polypeptide chains in proteins of known structure. Rather than exhaustively calculating the compactness of all possible units, our procedure uses a screening algorithm to find discontinuous regions that are potentially compact. Precise calculations of compactness are restricted only to units in these regions. With our procedure, compactness can be used to discover discontinuous domains with virtually any number of disjoint peptides. Small, single-domain proteins may contain several compact regions: thus, compact regions do not always correspond to folding domains. Because a domain is an independent folding unit and should contain a hydrophobic core, compact units were further examined for the presence of hydrophobic clusters (Zehfus MH, 1995, *Protein Sci 4*:1188–1202). This added constraint limits the number of acceptable units and helps greatly in the location of the true structural domains. The larger hydrophobically stabilized compact units correspond to domains, while the smaller units may correspond to folding intermediates.

Keywords: compact domains; compact units; discontinuous domains; domains; folding units; stabilized modules

Ever since the first protein structure was elucidated by X-ray crystallography, it has been possible to locate distinct, spatially separable regions within many proteins that are thought to correspond to structural domains (Wetlaufer, 1973). While many domain-finding algorithms were developed in the late 1970s and early 1980s (Liljas & Rossman, 1974; Crippen, 1978; Rose, 1979; Lesk & Rose, 1981; Rashin, 1981; Wodak & Janin, 1981; Go, 1983; Janin & Wodak, 1983; Kikuchi et al., 1988), there has been little activity in this area until recently. In the past two years at least five new methods for finding domains have been reported in the literature (Zehfus, 1994; Islam et al., 1995; Siddiqui & Barton, 1995; Sowdhamini & Blundell, 1995; Swindells, 1995a). Although most of these methods are logical extensions of principles and calculations used in earlier work, the emphasis has shifted toward faster, more flexible implementations. Faster domain-finding techniques are needed so the great number of proteins that now reside in the protein data bank can be analyzed, while more flexible techniques are being developed so discontinuous domains containing more that a single peptide chain can be found.

Each method has some potential difficulty. Several rely on Cα information alone, and ignore side-chain packing information (Islam et al., 1995; Sowdhamini & Blundell, 1995). Other methods assemble domains only from pieces of regular secondary structure, and do not recognize the presence of loops or more irregular structures in their domains (Sowdhamini & Blundell, 1995; Swindells, 1995a). Two of the techniques assemble discontinuous domains only from continuous domains, ignoring the possibility that the individual parts of a discontinuous domain may not be structurally independent (Islam et al., 1995; Siddiqui & Barton, 1995). Finally, most of these methods use iterative, contiguous binary divisions of the peptide chain to define their units, rather than allowing the N- and C-termini to be trimmed to optimize each domain individually (Islam et al., 1995; Siddiqui & Barton, 1995).

We have used compactness to define domains (Zehfus & Rose, 1986; Zehfus, 1987, 1994). This approach avoids the above problems because it uses both main-chain and side-chain atoms in its calculations, does not require any knowledge of protein secondary structure, does not assemble discontinuous units from continuous units, and optimizes the definition of each domain by adjusting N- and C-termini for best fit. The disadvantage of this method is that it is computationally intensive; a compactness calculation requires the determination of both the solvent accessible surface area and volume for each unit examined.

A 100-residue protein contains a few thousand continuous units, and millions of binary discontinuous units. This many compactness calculations can be exhaustively calculated in a reasonable length of time using a fast computer algorithm (Zehfus, 1993). The number of trinary (three peptides), and quadrinary (four peptides)

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

discontinuous units is expected to be in the billions and trillions, respectively, and so cannot be exhaustively calculated. This problem has been circumvented here using a simple screening parameter to quickly locate potentially compact regions. Once these regions have been found, then the more time consuming compactness calculation is used only on the units in that region, to determine the actual compactness of the domain. This greatly increases the efficiency of the compactness approach, making it applicable to the discovery of discontinuous units with any number of disjoint peptides.

The compactness analysis reveals that all proteins contain many compact regions. Regions identified only by their compactness are called *compact units*. These compact units are so numerous that they clearly cannot all be domains. Because compactness alone does not determine the presence of a domain, another property must be used to differentiate between compact regions and actual domains. If a domain is a piece of a protein that can fold independently, then it should contain a hydrophobic core large enough to stabilize its own structure. Because it was recently shown that hydrophobic clusters can be found using an adaptation of compactness theory (Zehfus, 1995), the compact units may be analyzed for the presence of a hydrophobic cluster within them. This simple filter removes roughly 75% of the compact units and allows the remaining units to be easily organized into closely related groups. The result is that only a few compact domains containing hydrophobic clusters are located in each protein. These structures are referred to here as *stabilized modules*. Although a few stabilized modules are large enough to correspond to distinct structural domains, most of the modules are smaller and may correspond to folding intermediates in a hydrophobic collapse folding pathway.

A

16-40 17-41 18-42 19-43 20-44
16-39 17-40 18-41 19-42 20-43 21-44
16-38 17-39 18-40 19-41 20-42 21-43 22-44
16-37 17-38 18-39 19-40 20-41 21-42 22-43 23-44
16-36 17-37 18-38 19-39 20-40 21-41 22-42 23-43 24-44
17-36 18-37 19-38 20-39 21-40 22-41 23-42 24-43
18-36 19-37 20-38 21-39 22-40 23-41 24-42
19-36 20-37 21-38 22-39 23-40 24-41
20-36 21-37 22-38 23-39 24-40

B

16-40 18-42 20-44
16-38 18-40 20-42 22-44
16-36 18-38 20-40 22-42 24-44
18-36 20-38 22-40 24-42
20-36 22-38 24-40

C

18-40 19-41 20-42
18-39 19-40 20-41 21-42
18-38 19-39 20-40 21-41 22-42
19-38 20-39 21-40 22-41
20-38 21-39 22-40

Fig. 1. Different peptides to be examined around a target peptide containing residues 20–40. (**A**) Full region, (**B**) region at lower resolution, (**C**) smaller region at high resolution.

Results and discussion

Overview of unit discovery method

A 100-residue protein contains about 5000 continuous units. If discontinuous units are generated by simply combining all possible continuous units, a protein this size would have millions of ternary (three peptide) units and billions of quadrinary (four peptide) units. Instead of trying exhaustively to calculate the compactness of each of these units, it is much more efficient to use a screening parameter to locate units that seem compact, and perform the compactness calculation only on these units. Because compact units are close to spherical in shape (Zehfus & Rose, 1986), spheres of varying sizes are moved through the protein's coordinates to find sets of residues that fit well within the spheres. These sets of residues are then used as seeds for further processing.

Once a seed unit is chosen, the region near that seed must be searched to find the most compact unit in that region. In previous work on unitary and binary domains, the region around a peptide was varied by ±4 residues at each end, with a total change in size of ±4 residues. This region is shown in Figure 1A. It can be seen that this region includes 61 slightly different peptides. Although this sounds like a small number of peptides to examine, when four of these regions are combined in a quadrinary unit, there are $61 \times 61 \times 61 \times 61$, or more than 13.8 million combinations! Rather than evaluating the Z of all these units, a two-pass procedure, one over the entire region at low resolution (Fig. 1B), and a second over a smaller region at higher resolution (Fig. 1C) is used to reduce the number of Z calculations performed. This two-stage process reduces the number of calculations needed to locate a minimum by at least 94%, and further speeds the discovery of compact units.

Once a list of compact units is developed, the protein is subjected to an independent search for hydrophobic clusters (Zehfus, 1995). The list of compact units is then analyzed and units that do not contain hydrophobic clusters are removed. By comparing the underlying clusters the remaining units can be easily grouped, and redundant structures eliminated. The resulting structures are the final hydrophobically stabilized modules.

Efficacy of the screening procedure

In the original algorithms used to find continuous and binary discontinuous compact domains the compactness of all possible units was calculated, so it was impossible to miss a potential compact unit (Zehfus, 1993, 1994). In the seeded search algorithm the compactness of a limited set of peptides is determined, so it is possible that some truly compact units might be missed.

To try to avoid this, the parameters used in the seeding procedure were optimized using ribonuclease A as a test system. In this optimization the discontinuous binary units of ribonuclease were determined both exhaustively and using the seeded search method. The seeded search parameters were then adjusted until all compact units were discovered.

The search parameters optimized in this process were: step size for sphere movement, step size of sphere radius, and the threshold for number of Cαs found in the sphere to qualify as seed structure. The optimal values for step of sphere movement and sphere radius

are 1.0 Å and 0.5 Å, respectively. If either of these parameters is set to a higher value, some units are lost. If they are set lower, the search is performed at a finer resolution, but no additional units are found.

The threshold for number of Cαs needed in a sphere to qualify as a seed structure is set at two standard deviations above the mean value. The threshold value represents a tradeoff. If the threshold is set lower, there is a greater certainty of finding all compact units, but the number of seeds that result in non-compact units increases greatly, as does overall computation time. If the threshold is set higher, then fewer seed structures are evaluated, and the computation time is lowered; however, the higher threshold increases the chances that a true compact unit will be missed.

The appropriateness of these parameters was tested by comparing the results obtained when the binary units of five additional proteins were determined using both seeded and exhaustive search procedures. The proteins used were cytochrome *c* (3cyt), interleukin (4ilb), myoglobin (4mbn), hen egg white lysozyme (6lyz), and pancreatic trypsin inhibitor (6pti). In this set of proteins 98 compact clusters are found using exhaustive evaluation. When evaluated using the seeding method, eight additional units are found that are not true compact units, and eight true compact units are missed. Seven out of the eight missed units are only moderatly compact ($Z \geq 1.50$), so the seed procedure found virtually all significant units. An additional 14 units are missed by the seeding procedure that are either borderline units and not considered significant, or are variants of better units that are found correctly.

Most errors in the seed method occur in units of marginal compactness. Because tri- and quad-discontinuous units are generally more compact than uni- or bi-units, both the number of marginal cases, and the overall error rate is expected to be lower for tri- and quad-units.

The screening procedure is efficient, only 15% of the seeded structures are rejected because they are non-compact. One area that needs further attention, however, is the removal of seeds containing small peptides; 45% of the seeds lead to structures that are rejected because a peptide within it is too small.

Efficacy of screening compact units for enclosed hydrophobic clusters

Most proteins contain many compact regions. Ribonuclease, for example, contains 1 unitary, 10 binary, 17 trinary, and 4 quadrinary compact units. Because a protein of this size should contain only one or two domains, clearly the bulk of these compact units are not domains, and compactness alone does not determine the presence of a domain.

If a domain is a piece of a protein that can fold independently, then it should contain a hydrophobic core large enough to stabilize its own structure. The compact units were therefor examined to see if they contained hydrophobic clusters. To do this the protein was subjected to an independent hydrophobic cluster analysis (Zehfus, 1995). Only those compact units that contained a hydrophobic cluster within them were retained for further investigation. When a compact region is stabilized by a hydrophobic cluster it is called a *stabilized module*. In ribonuclease there are only two binary, eight trinary, and two quadrinary stabilized modules. Further, when these modules are organized by their underlying hydrophobic clusters, it becomes obvious that there are only two groups of similar units; one set of binary modules containing residues 23–46 and 82–100, and a set of trinary modules containing roughly residues 4–14, 45–82, and 102–124.

Table 1. *Compact units, stabilized modules, and structurally distinct stabilized modules found in 14 surveyed proteins*

Peptides in unit	Compact units	Stabilized modules	Distinct structures
1	117	22	18
2	237	47	17
3	234	75	23
4	54	20	11

Table 1 summarizes the stabilized modules and unique structures found in the set of 14 proteins analyzed here. There is a total of 642 compact units in these proteins, but only about a quarter (162) correspond to stabilized modules; thus, the requirement for a hydrophobic cluster is relatively stringent. When these stabilized modules are further examined to eliminate similar units, only about 20 distinct 1-, 2-, and 3-peptide modules and about 10 distinct 4-peptide modules remain.

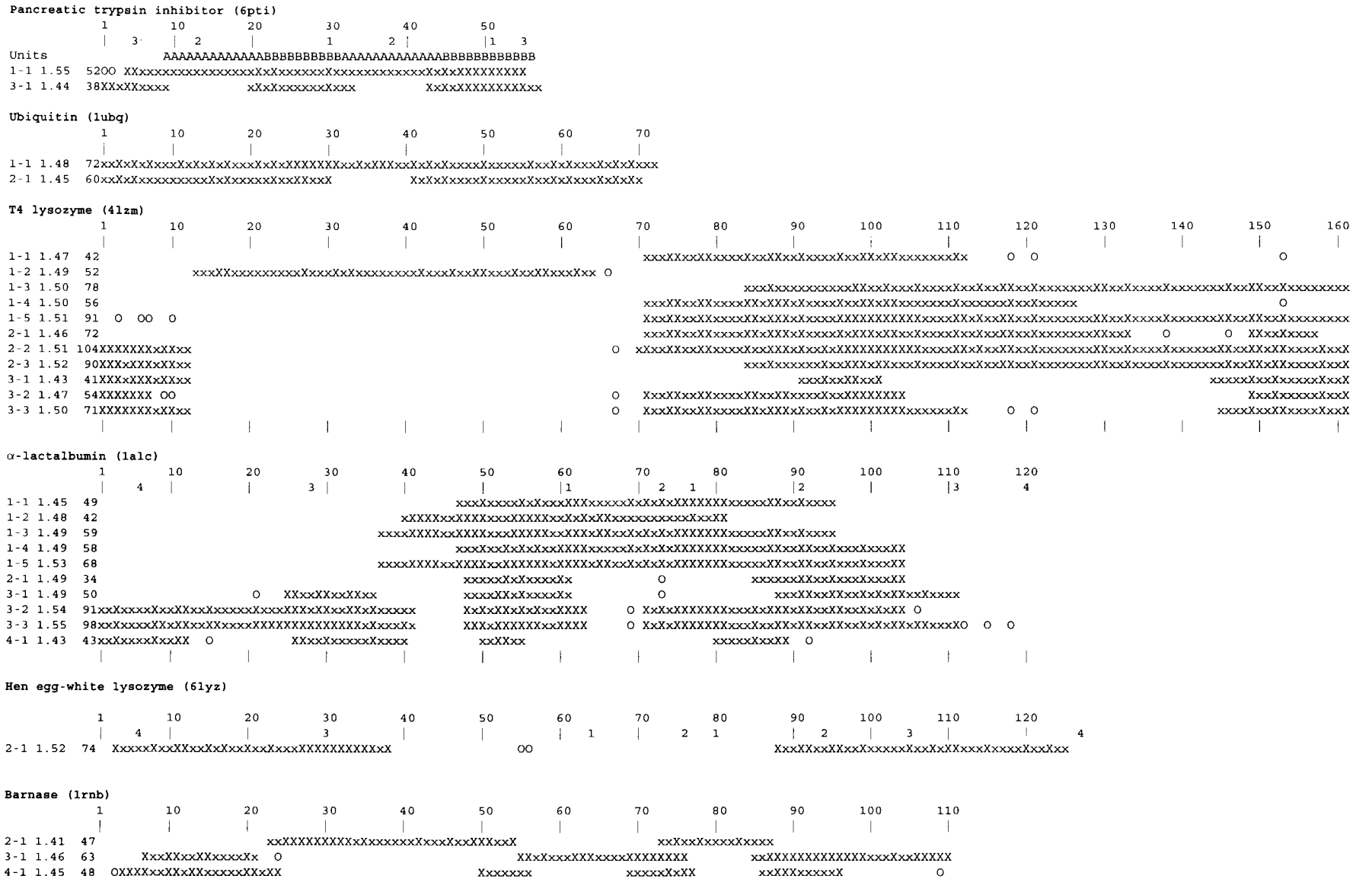
Results of individual proteins

Figure 2 lists the distinct stabilized modules found in the proteins studied here, while Figure 3 displays the stabilized modules from four arbitrarily chosen example proteins. Some of these proteins have a very simple anatomy, while others are quite complex. Pancreatic trypsin inhibitor, for instance, has only two stabilized modules, one that corresponds to the entire protein and a second trinary unit that comprises roughly 2/3 of the molecule. This unit corresponds closely to the 1-9 + 20-33 + 42-58 PαPγ peptide shown by Staley and Kim (Staley & Kim, 1990) to have native-like structure. This unit had not been identified by compactness previously because it contains three disjoint peptides. A compact unit corresponding to the other third of the molecule can be found, but it is not considered a stabilized module because it does not contain a hydrophobic cluster.

The stabilized modules of ubiquitin are also very simple. Again, one module corresponds to the entire protein, while the second is essentially the same unit, but with a short piece of the C-terminus and a neighboring loop removed to optimize compactness.

By eye T4 lysozyme is clearly a two-domain protein. The N-terminal domain contains both α and β structure, while the C-terminal domain is discontinuous and primarily helical in nature. This two domain structure is captured perfectly in the two major stabilized modules 1-2 and 2-2. The C-terminal helical domain contains many smaller alternative stabilized clusters. The exact meaning of these alternative modules is not clear. One likely explanation is that these alternate units simply represent different ways of repacking the domain core using different sets of structural units. Although it is possible that these alternate units could correspond to intermediates in a folding pathway, little correlation is seen between these units and the hydrogen exchange data of an early folding intermediate identified by Lu and Dahlquist (1992).

The structurally homologous α-lactalbumin and hen egg white lysozyme present an interesting contrast with each other. These structures are visually similar, but hen egg white lysozyme contains a single discontinuous stabilized module corresponding a union of the N- and C-termini of the protein, while α-lactalbumin has several


Pancreatic trypsin inhibitor (6pti)
Units
1-1 1.55 5200
3-1 1.44 38
Ubiquitin (1ubq)
1-1 1.48 72
2-1 1.45 60
T4 lysozyme (41zm)
1-1 1.47 42
1-2 1.49 52
1-3 1.50 78
1-4 1.50 56
1-5 1.51 91
2-1 1.46 72
2-2 1.51 104
2-3 1.52 90
3-1 1.43 41
3-2 1.47 54
3-3 1.50 71
α-lactalbumin (1alc)
1-1 1.45 49
1-2 1.48 42
1-3 1.49 59
1-4 1.49 58
1-5 1.53 68
2-1 1.49 34
3-1 1.49 50
3-2 1.54 91
3-3 1.55 98
4-1 1.43 43
Hen egg-white lysozyme (6lyz)
2-1 1.52 74
Barnase (1rnb)
2-1 1.41 47
3-1 1.46 63
4-1 1.45 48

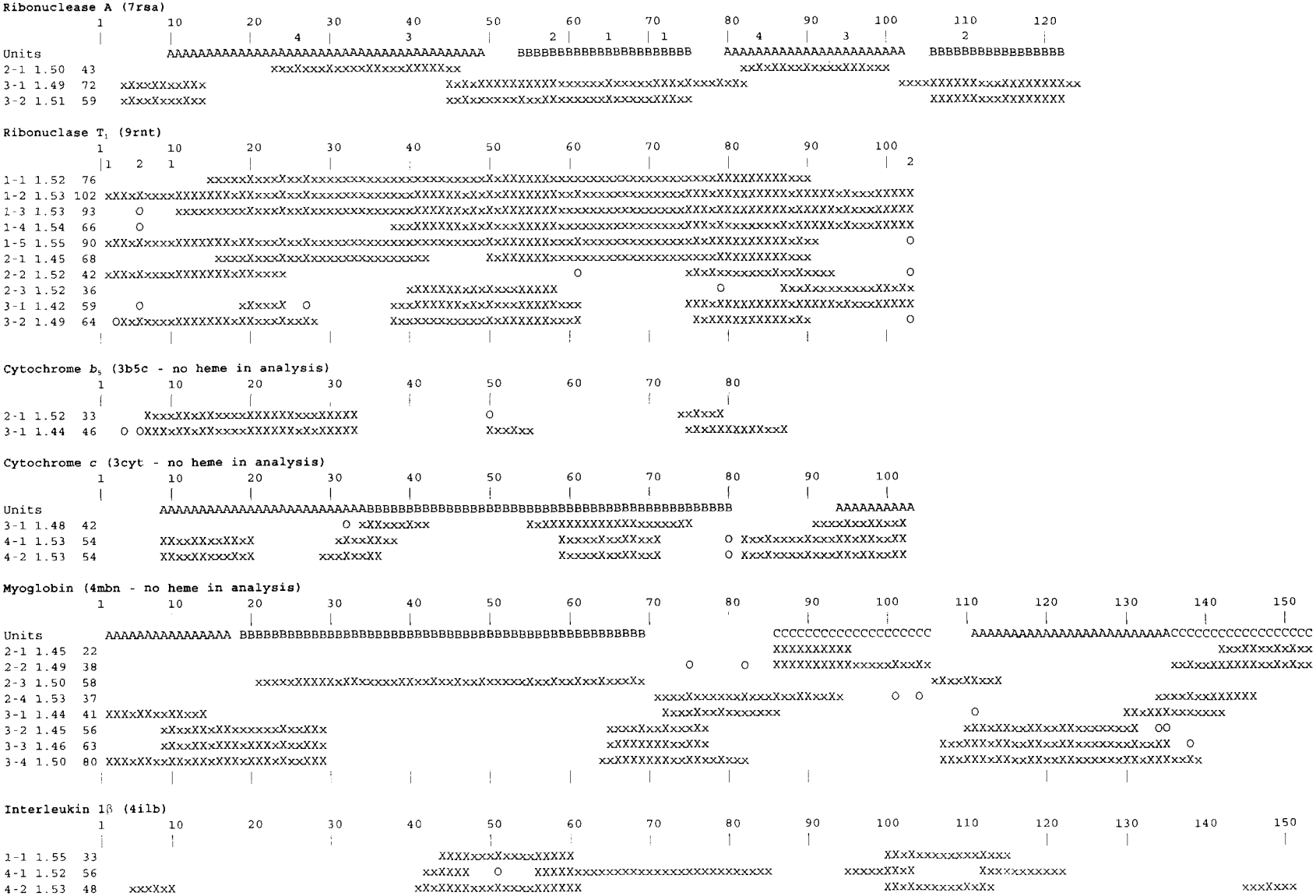


Fig. 2. *Continues.*

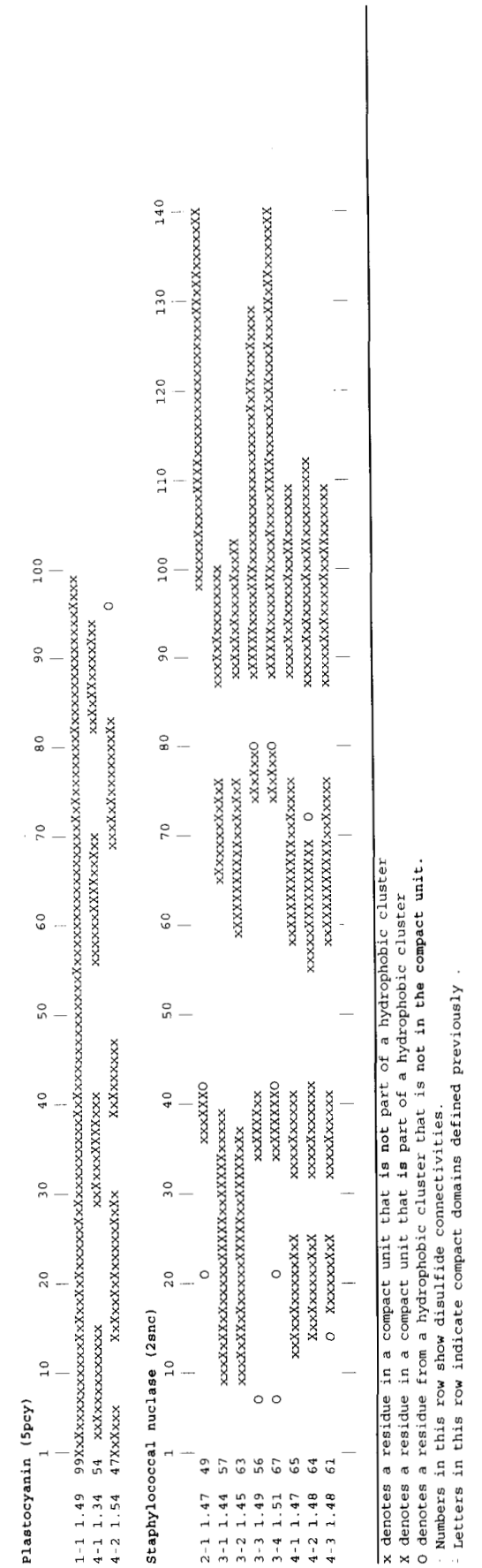


Fig. 2. Individual compact, hydrophobically stabilized domains and modules (stabilized modules) found in 14 proteins.

different stabilized modules, most centered on the continuous region around the calcium binding loop between residues 80 and 90. This dichotomy occurs because the region in α-lactalbumin equivalent to the discontinuous lysozyme domain is not compact in α-lactalbumin, while the region in lysozyme equivalent to the α-lactalbumin continuous domain does not contain a hydrophobic cluster in lysozyme. The discontinuous helical domain of hen egg white lysozyme had been detected as a folding intermediate in both water (Miranker et al., 1991) and TFE (Buck et al., 1993). The continuous module from α-lactalbumin corresponds closely to a region that is important in the transition between a molten globule and a native state that occurs when this protein binds calcium (Wu et al., 1996).

Although barnase contains three stabilized modules, the trinary and quadrinary modules are essentially the same unit, so the quadrinary unit may be regarded as redundant. Fersht's group has studied this molecule intensely (Fersht et al., 1992; Matouschek et al., 1992a, 1992b; Serrano et al., 1992a, 1992b, 1992c) and has developed a model where two hydrophobic core regions are thought to fold quickly, while a third hydrophobic core is folded at a slower rate. The large trinary stabilized module corresponds quite well with the fast folding $core_1$ and $core_3$, while the smaller binary unit is equivalent to the slower folding $core_2$.

Ribonuclease A contains a cleft that delineates the protein into two distinct regions, identified as separate domains by Wetlaufer in 1973 (Wetlaufer, 1973). Because one of these domains (1–10 + 52–76 + 105–124) is trinary in nature, it could not be properly identified using compactness until now. The analysis actually finds two very similar trinary stabilized modules, but the 3-1 module (4–14 + 45–82 + 102–124) is preferred because it is both larger and more compact. The remainder of the protein is in a binary stabilized module (23–46 and 82–100) that is complementary to the trinary unit. Both the bi- and the tri-modules contain disulfides that hold them together. In the binary unit both disulfides bridge the discontinuous peptides together, while in the trinary unit one disulfide is located within a peptide, and the second joins two of the three peptides together.

Ribonuclease T_1 contains several stabilized modules. Many of these units correspond to the entire protein with small pieces removed to optimize compactness (1-1, 1-2, 1-3, 1-5, and 2-1). The remaining units are highly overlapped and no set of mutually exclusive regions can be identified; thus, this protein appears to contain a single folding domain. Hydrogen exchange data have shown that amino acids in the middle and N- and C-termini are associated with a fast-folding native-like intermediate (Mullins et al., 1993), and these regions are observed in the trinary stabilized modules found here.

The proteins cytochrome b_5, cytochrome c, and myoglobin contain heme groups. The analysis method works only with amino acids and ignores prosthetic groups, so the heme plays no role in the stabilized modules of these proteins. Cytochrome b_5 contains two similar units, one binary, and the second trinary. The binary unit corresponds well with the region found by Moore and Lecomte (Moore & Lecomte, 1990) to have structure in the apo-cytochrome molecule. The trinary unit is composed of the binary unit plus a very small β-strand. While this addition greatly enhances the compactness of the binary unit, this peptide is just at the minimum allowed size for a peptide in a discontinuous unit, so its structural significance is questionable. Further, no experimental evidence for this trinary unit is seen in Moore and Lecomte's data.

Cytochrome c contains three modules: one trinary, and two essentially equivalent quadrinary units. The quadrinary units corre-

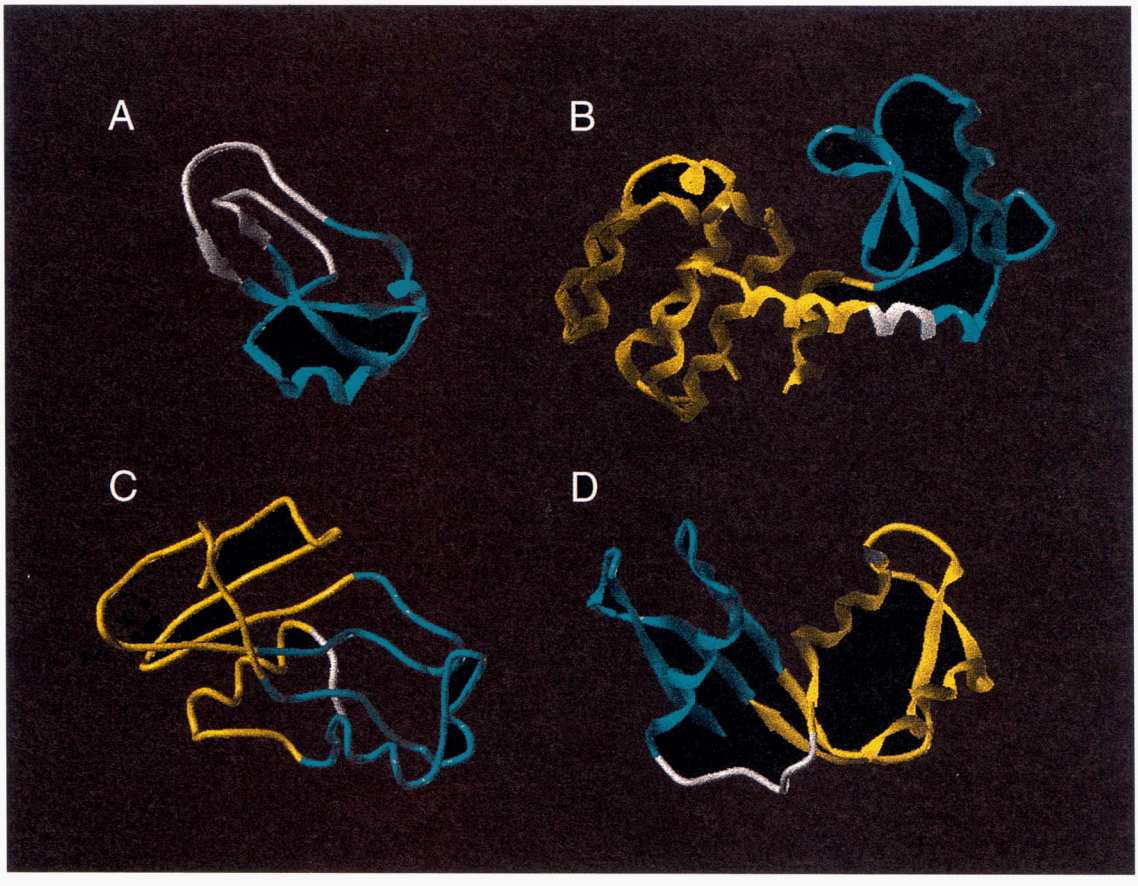


Fig. 3. Stabilized domain and modules from four example proteins. (**A**) BPTI, blue: unit 3-1, residues 1–9 + 20–33 + 43–57, white: remainder. (**B**) T4 lysozyme, blue: unit 1-2, residues 13–64, yellow: unit 2-2, residues 1–12 + 70–161, white: remainder. (**C**) Barnase, blue: unit 2-1, residues 23–54 + 73–87, yellow: unit 3-1, residues 7–21 + 55–76 + 85–100, white: remainder. (**D**) Ribonuclease A, blue: unit 2-1, residues 23–46 + 82–100, yellow: unit 3-1, residues 4–14 + 45–82 + 102–124, white: remainder.

spond to the addition of an N-terminal helix to the trinary unit. Roder et al. (Roder et al., 1988) has used hydrogen exchange in cytochrome *c* to identify the N- and C-terminal helices as important early folding intermediates. Jeng et al. (Jeng et al., 1990; Jeng & Englander, 1991) have studied this protein in a nonnative compact state and further implicated the 60s helix as important in folding. All three of these structures are observed in the quadrinary stabilized module. An entire folding and unfolding pathway for this protein has recently been proposed by Bai et al. (Bai et al., 1995) by studying proton exchange as a function of guanidinium concentration. Although there are similarities between their proposed folding structures and the stabilized units found here, there are also some differences. This is most likely due to the fact that the experimental studies are done in the presence of covalently bound heme, while the analysis performed here does not take the heme group into account.

Myoglobin contains several binary and trinary stabilized modules. The largest of the trinary units corresponds quite well with a region observed by Hughson et. al. (Hughson et al., 1990) to have structure in apo-myoglobin. The N-terminal peptide of the largest binary unit also corresponds to a tertiary cluster observed by Coco and Lecomte in apo-myoglobin (Cocco & Lecomte, 1990).

Interleukin 1β is an irregular beta barrel protein. The extremely discontinuous nature of this beta barrel has made it difficult to analyze using compactness. The binary stabilized module found here represents two β-turn-β structures that stabilize half of the protein's β-barrel core. The 4-1 unit extends this structure along one side of the barrel, while the 4-2 structure completes the basic barrel. Interestingly, the 4-1 stabilized module contains a large segment between residues 60 and 85 that does not seem to be stabilized by a hydrophobic core, yet this unit corresponds well with a region observed by Varley et al. (1993) to be an early folding region in this protein.

Plastocyanin is a small β-barrel protein that contains three stabilized modules. One module corresponds to the entire protein, while the other modules coincide with the two ends of the barrel. These two units bisect the protein's β-strands between them. It seems unlikely that a protein could have two independent domains at opposite ends of a small β-barrel; thus, this protein is problematic. A likely explanation of this problem is that the algorithm that finds the protein's hydrophobic clusters has incorrectly divided a large somewhat non-compact cluster into two smaller, but more compact ones (Zehfus, 1995). This incorrect division is then used to identify two separate stabilized modules. It appears that the algorithm used to find hydrophobic clusters needs further refinement to remove this artifact.

The 2-1 binary discontinuous unit of staphylococcal nuclease contains one large peptide and a much smaller six-residue fragment. Although the significance of such small peptide fragments is questionable, this fragment may play an important role here.

Experiments on the isolated large peptide of this unit show only marginal, nascent structure in aqueous solution (Maciejewski, 1996). Because the continuous unit itself does not correspond to a stabilized module because it does not contain a complete hydrophobic cluster, it appears that the hydrophobic residues contained in the 36–41 fragment may be critical to the overall folding of this unit.

The remaining units of staphylococcal nuclease are very similar and seem to indicate the importance of a trinary unit in this protein's structure.

Discussion

Picking structural units in a protein based on compactness is evolving in sophistication. Originally, compact units were chosen only on the basis of compactness, and the "best" units in a protein were those that were most compact (Zehfus, 1987). Later, when the approach was expanded to include binary discontinuous units, it was recognized that some very compact units might not be good structural units, while other less compact structures made better structural sense. At that time the best set of units for a protein were selected on the basis of mutual exclusivity and best average compactness for the set of units as a whole (Zehfus, 1994). In this work it is recognized that one of the main forces stabilizing protein structure is hydrophobic in nature, and units are selected that are both compact and contain a hydrophobic cluster within them. This makes the units found here different than the compact units found previously. Where available, Figure 2 includes information on previously published compact domain assignments.

What is the structural significance of the stabilized modules found here? Several of the larger stabilized modules (2-1 and 3-1 in ribonuclease and 1-2 and 2-2 in T4 lysozyme) clearly correspond to classical domains. On the other hand, several proteins that are thought to contain a single domain are found to contain a number of stabilized modules, often spanning similar regions of the protein. In these cases it appears that the region is simply being redefined in alternate ways to try to optimize its compactness. Further refinement of the compact unit finding algorithm is needed in these cases to more clearly identify the optimum unit.

If the larger stabilized modules correspond to independent structural domains, then it seems logical that the smaller stabilized modules should correspond to pieces of structure that have a propensity to fold into their respective structure. This idea is reinforced by the fact that all stabilized modules contain hydrophobic clusters, and these clusters have been shown to correlate with early folding intermediates (Zehfus, 1995). It seems likely, then, that stabilized modules may correspond to protein folding intermediates. If so, how do these units fit into current models of protein folding? The framework model of protein folding proposes that elements of secondary structure form first, and then coalesce to form larger structures, while the hydrophobic collapse theory holds that hydrophobic regions collapse first, followed by the formation of secondary and tertiary structure.

Stabilized modules, with their hydrophobic core and compact structure, are well suited to be regions of the protein prone to hydrophobic collapse, and should be considered as possible intermediates in the hydrophobic collapse folding pathway. If stabilized modules do represent intermediates in protein folding, then they should be observable in protein folding experiments. The proteins studied here were chosen because their folding pathways have been studied. Good correlation is seen between stabilized modules and folding intermediates in several cases, but in others, little correlation is observed. These cases of poor correlation, however, do not necessarily mean that stabilized modules do not correspond to folding intermediates.

This analysis has tried to correlate a large number of different experimental techniques thought to probe early events in protein folding with stabilized modules. Some of these methods monitor kinetic pathways, while others rely on the identification of thermodynamically stable intermediates under a wide variety of native and non-native conditions. At the present time it is not clear how these techniques correlate with each other, let alone how they correlate with the true folding pathway (Clarke & Fersht 1996). The final result is that a perfect correlation with experimental data cannot be expected.

Comparing the results obtained here with other domain finding algorithms (Islam et al., 1995; Siddiqui & Barton, 1995; Sowdhamini & Blundell, 1995; Swindells, 1995a) on a protein-for-protein basis is difficult because there is little overlap between sets of analyzed proteins. This poor overlap is due to a difference in focus. Most other methods emphasize large proteins containing hundreds of residues, and look primarily for large, structurally independent domains. Although some of the larger stabilized modules correspond to this kind of independent domain, the bulk of the structures found here are smaller, and the other methods are simply not designed to find such units.

One area where stabilized clusters differ from large independent domains is in the number of discontinuous units discovered. In their procedure Siddiqui and Barton (Siddiqui & Barton, 1995) found 190 uni-, 41 bi-, 1 tri-, and 1 quad-domains. Similarly, Islam et al. (Islam et al., 1995) found 152 uni-, 2 bi-, and only 3 tri-domains. It appears that the bulk of the discovered large domains are composed of a single peptide, while discontinuous units, especially those containing more than two peptides, are much rarer. Here the distribution of uni-, bi-, and tri-units are approximately the same, and it is only at the quadrinary level where a significant decrease in population is found (Table 1).

Methods

The proteins analyzed here are: α-lactalbumin (1alc) (Acharya et al., 1989), barnase (1rnb) (Baudet & Janin, 1991), ubiquitin (1ubq) (Vijay-Kumar et al., 1987), cytochrome b_5 (3b5c) (Mathews et al., 1971), cytochrome c (3cyt) (Takano & Dickerson, 1980), interleukin 1β (4ilb) (Veerapandian et al., 1992), T4 lysozyme (4lzm) (Bell et al., 1991), myoglobin (4mbn) (Takano, 1984), hen egg-white lysozyme (6lyz) (Diamond, 1974), pancreatic trypsin inhibitor (6pti) (Wlodawer et al., 1987), ribonuclease A (7rsa) (Wlodawer et al., 1988), ribonuclease T_1 (9rnt) (Martinez-Oyanedel et al., 1991), plastocyanin (5pcy) (Guss et al., 1986), and staphylococcal nuclease (2snc) (Loll & Lattman, 1989). All coordinates were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977). As in previous work, the compactness analysis is restricted to amino acids only, ions or prosthetic groups are not included in any calculations.

Screening method

To improve the speed of the screening function, only the positions of the Cα atoms are used. To locate locally compact regions in a protein, a series of different-sized spheres are passed through the protein's coordinates, and a record is kept of any groups of Cα atoms that fit well within the spheres. Although several different

sphere sizes and movement steps were tried, the final parameters used here vary the sphere radius from the average radius of the protein down to 2.5 Å in steps of 0.5 Å, and the sphere is moved in 1 Å steps in the *X*, *Y*, and *Z* directions. This set of parameters was chosen because it would find all the compact binary units in ribonuclease, while smaller changes would provide no additional units (see Results and discussion).

In this process only pieces containing four or more contiguous Cα atoms are used, and any smaller pieces of structure are discarded. As each sphere is passed through the protein, a running total is kept of the number of atoms in each binary, trinary, and quadrinary set of peptides that fit within the sphere. From this data the average and standard deviation of the number of Cαs in bi-, tri- and quad-units is determined for each sphere size. This information is then used to identify units that have a significantly higher than average number of Cαs packed into a sphere. Using ribonuclease it was determined empirically that if a threshold of two standard deviations above the mean is used as a cutoff, all the compact binary discontinuous units in this protein are properly identified (see Results and discussion). This level was therefore used for the selection of compact seeds in unknown proteins.

In ribonuclease this threshold identifies 11,185 bi-, 18,410 tri-, and 11,743 quad-units as possible seeds for discontinuous domains. Many of these units are very closely related, differing from each other by only a few residues. As a result, potential seeds are next grouped together, and a single "average" unit is picked as a seed for each region. In this procedure the list of potential units is first scanned to find the one unit that has the largest set of closely related neighboring units, i.e., units that vary by ± 2 or less in any peptide parameter (N-terminal, residue, C-terminal residue, or number of residues). These units are grouped together and a single average set of parameters calculated to represent this set of units. The units contributing to this set are eliminated from the list, and the process is continued to identify the next most common grouping.

These groupings are then inspected to see if they can be further merged together. Here, the average parameters for each group are compared to each other. Groups are merged only if the maximum difference between the two sets is less than a given threshold. This threshold starts with a maximum difference of ± 1 and then increased until the maximum allowed difference in any parameter is ± 4. Naturally, as groups are merged, new average definitions for the group are recalculated based on the set of units within the grouping. The final set of average parameters for each group are then used as seeds for the next step of processing.

In ribonuclease this trims the list of potential units down to 483 bi-, 697 tri-, and 475 quad-seeds to be used in the domain finding process.

Evaluation of screened peptides

Once a list of seed units has been determined, the most compact unit near that seed must be found. As shown in Results and discussion, the region around each peptide normally searched contains 61 slightly different peptides (Fig. 1A). When this many different definitions for one peptide are combined with multiple definitions for two or three other peptides in a discontinuous unit, the number of combinations can be in the millions, and an exhaustive search is not practical. Instead, a two-step search is used that lowers the number of units examined by more than 90%.

In the first step of this search, the entire region around each peptide is searched at low resolution, as shown in Figure 1B. Because this region includes only 19 peptides, when it is combined with similar regions from other peptides in the unit, the total number of combinations is much lower. In the second step of the search the region closest to the seed is examined at the high resolution, as shown in Figure 1C. Again, this region contains only 19 members, and the number of combinations is relatively small.

If the most compact unit found in these two searches is the original seed, the search is ended because the most compact unit has been found. If the most compact unit does not correspond to the seed, then the position of the most compact unit is used as a new seed, and the search is re-initiated at this new site. The search continues until a minimum is found, or one of the peptides in the seed unit contains less than five residues.

Once well-defined compact units are found for each of the seed structures, the list of compact units is processed using standard criteria (Zehfus, 1994). These criteria are: all peptides in each unit must have six or more residues, the gap between two peptides in a unit must be six or greater, and the Z of each unit must be 1.55 or lower. Units that pass the above test are called *compact units*.

Using hydrophobic clusters to filter potential domains

Hydrophobic clusters are determined independently using the method of Zehfus (Zehfus, 1995). Each compact unit is then tested to see if it contains a hydrophobic cluster within it. Two simple rules are used in the determination. First, at least 90% of a hydrophobic cluster's residues must be contained within the unit; and second, if the unit is discontinuous, then each disjoint peptide within it must contain some residues from the hydrophobic cluster. Units that pass this test are called *stabilized modules*.

Hydrophobic clusters are also used to group together similar stabilized modules. After each protein is analyzed, all stabilized modules having the same underlying clusters are identified. If the same set of hydrophobic clusters stabilizes multiple modules, then a single best unit is identified and the other units are eliminated. The factors of compactness, size, fit with the underlying clusters, and number of peptides in the unit are the criteria used to select the "best" stabilized module.

Correction for size dependancy in tri- and quad-units

In previous work it was noticed the *Z* function of continuous and binary discontinuous units was dependent on size, and an empirical size correction equation was derived to remove this factor (Zehfus & Rose, 1986; Zehfus, 1994). This is also true for trinary and quadrinary discontinuous units. The size dependancy was detected using the proteins α-lactalbumin (1alc), cytochrome *c* (3cyt), dihydrofolate reductase (4dfr), hen egg-white lysozyme (6lyz), T4 lysozyme (4lzm), barnase (1rnb), ribonuclease (7rsa), Staphylococcal nuclease (2snc), interleukin 1β (4ilb), myoglobin (4mbn), papain (9pap), Bence-Jones immunoglobulin (1rei), subtilisin (1sbt), and superoxide dismutase (2sod). Here 10,000 trinary and quadrinary discontinuous units containing 15, 20, 25, ... 60 residues were randomly chosen from each protein. The average and standard deviation of these random samples was then determined. As done previously (Zehfus & Rose, 1986), an empirical correction factor was then determined that would make the mean minus one standard deviation linear between 5 and 40 residues. No correction factor was used for units containing more than 40 residues. The correction factors for tri- and quad-units are given below:

Correction factor for ternary units

$$= 0.534e^{(-0.0581 \times \text{number of residues})} + 0.948$$

Correction factor for quadrinary units

$$= 0.413e^{(-0.0616 \times \text{number of residues})} + 0.965$$

Acknowledgment

This work was supported by grant number GM46664 from the National Institutes of Health.

References

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

Bai Y, Sosnick TR, Mayne L, Englander SW. 1995. Protein folding intermediates: Native-state hydrogen exchange. *Science 269*:192–197.

Baudet S, Janin J. 1991. Crystal structure of a barnase-(dGpC) complex at 1.9 A 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.

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.

Clark J, Fersht AR. 1996. An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. *Folding Design 1*:243–254.

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

Crippen GM. 1978. The tree structural organization of domains in globular proteins. *J Mol Biol 126*:315–332.

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

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.

Go M. 1983. Modular structural units, exons, and function in chicken lysozyme. *Proc Natl Acad Sci USA 80*:1964–1968.

Guss JM, Harrowell PR, Murata M, Norris VA, Freeman HC. 1986. Crystal structure analyses of reduced (Cu^{I}) poplar plastocyanin at six pH values. *J Mol Biol 192*:361–387.

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

Islam SA, Luo J, Sternberg MJE. 1995. Identification and analysis of domains in proteins. *Protein Eng 8*:513–525.

Janin J, Wodak SJ. 1983. Structural domains in proteins and their role in the dynamics of protein function. *Prog Biophys Mol Biol 42*:21–78.

Jeng M-F, Englander SW. 1991. Stable submolecular folding units in a non-compact form of cytochrome *c*. *J Mol Biol 221*:1045–1061.

Jeng M-F, 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.

Kikuchi T, Nemethy G, Scheraga HA. 1988. Prediction of the location of structural domains in globular proteins. *J Protein Chem 7*:427–471.

Lesk AM, Rose GD. 1981. Folding units in globular proteins. *Proc Natl Acad Sci USA 78*:4304–4308.

Liljas A, Rossman MG. 1974. X-ray studies of protein interactions. *Annu Rev Biochem 43*:475–507.

Loll PJ, Lattman EE. 1989. The crystal structure of the ternary complex of staphylococcal nuclease Ca^{+2} and the inhibitor pdTp, refined at 1.65 Å. *Proteins Struct Funct Genet 5*:183–201.

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.

Maciejewski MW. 1996. Structural characterization of compact peptides from Staphylococcal nuclease by circular dischroism and nuclear magnetic resonance spectroscopy [dissertation]. Columbus: The Ohio State University.

Martinez-Oyanedel J, Choe HW, Heinemann U, Saenger W. 1991. Ribonuclease T1 with free recognition and catalytic site: Crystal structure analysis at 1.5 A resolution. *J Mol Biol 222*:335–352.

Mathews FS, Argos P, Levine M. 1971. The structure of cytochrome b_5 at 2.0 Å resolution. *Cold Spring Harb 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. 1990. Structural properties of apocytochrome b5: Presence of a stable native core. *Biochemistry 29*:1984–1989.

Mullins LS, Pace CN, Raushel FM. 1993. Investigation of ribonuclease T1 folding intermediates by hydrogen-deuterium amide exchange-two-dimensional NMR spectroscopy. *Biochemistry 32*:6152–6156.

Rashin AA. 1981. Locations of domains in globular proteins. *Nature 291*:85–86.

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

Rose GD. 1979. Hierarchic organization of domains in globular proteins. *J Mol Biol 134*:447–470.

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.

Siddiqui AS, Barton GJ. 1995. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci 4*:872–884.

Sowdhamini R, Blundell TL. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci 4*:506–520.

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

Swindells MB. 1995a. A procedure for detecting structural domains in proteins. *Protein Sci 4*:103–112.

Swindells MB. 1995b. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci 4*:93–102.

Takano T. 1984. Refinement of myoglobin and cytochrome *c*. In: *Methods and applications in crystallographic computing*. Oxford, UK: Clarendon 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.

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 A resolution. *J Mol Biol 194*:531–544.

Wetlaufer DB. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA 70*:697–701.

Wlodawer A., Walter J, Gilliland GL, Gallagher W, Woodward C. 1987. Structure of bovine 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.

Wodak SJ, Janin J. 1981. Location of structural domains in proteins. *Biochemistry 20*:6544–6552.

Wu LC, Schulman BA, Peng Z-Y, Kim PS. 1996. Disulfide determinats of calcium-induced packing in α-lactalbumin. *Biochemistry 35*:859–863.

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

Zehfus MH. 1993. Improved calculations of compactness and a reevaluation of continouous compact units. *Proteins Struct Funct Genet 16*:293–300.

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

Zehfus MH. 1995. Automatic recognition of hydrophobic clusters and their correlation with protein folding units. *Protein Sci 4*:1188–1202.

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