Hydrophobic folding units derived from dissimilar monomer structures and their interactions

CHUNG-JUNG TSAI¹ AND RUTH NUSSINOV^{2,3}

¹Laboratory of Mathematical Biology, NCI-FCRDC, Bldg. 469, Rm. 151, Frederick, Maryland 21702

²Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv 69978, Israel

³Laboratory of Mathematical Biology, SAIC, NCI-FCRDC, Bldg. 469, Rm. 151, Frederick, Maryland 21702

(RECEIVED July 10, 1996; ACCEPTED September 26, 1996)

Abstract

We have designed an automated procedure to cut a protein into compact hydrophobic folding units. The hydrophobic units are large enough to contain tertiary non-local interactions, reflecting potential nucleation sites during protein folding. The quality of a hydrophobic folding unit is evaluated by four criteria. The first two correspond to visual characterization of a structural domain, namely, compactness and extent of isolation. We use the definition of Zehfus and Rose (Zehfus MH, Rose GD, 1986, *Biochemistry* 25:335–340) to calculate the compactness of a cut protein unit. The isolation of a unit is based on the solvent accessible surface area (ASA) originally buried in the interior and exposed to the solvent after cutting. The third quantity is the hydrophobicity, equivalent to the fraction of the buried non-polar ASA with respect to the total non-polar ASA. The last criterion in the evaluation of a folding unit is the number of segments it includes. To conform with the rationale of obtaining hydrophobic units, which may relate to early folding events, the hydrophobic interactions are implicitly and explicitly applied in their generation and assessment. We follow Holm and Sander (Holm L, Sander C, 1994, *Proteins* 19:256–268) to reduce the multiple cutting-point problem to a one-dimensional search for all reasonable trial cuts. However, as here we focus on the hydrophobic cores, the contact matrix used to obtain the first non-trivial eigenvector contains only hydrophobic contacts, rather than all, hydrophobic and hydrophilic, interactions. This dataset of hydrophobic folding units, derived from structurally dissimilar single chain monomers, is particularly useful for investigations of the mechanism of protein folding. For cases where there are kinetic data, the one or more hydrophobic folding units generated for a protein correlate with the two or with the three-state folding process observed. We carry out extensive amino acid sequence order independent structural comparisons to generate a structurally non-redundant set of hydrophobic folding units for fold recognition and for statistical purposes.

Keywords: compactness; domain cutting; folding unit; hydrophobic core; protein folding

With the growing number of protein structures determined by X-ray crystallography and by nuclear magnetic resonance, it has become increasingly clear that the hydrophobic effect is the dominant driving force in protein folding (Dill, 1990). In this scenario, most of the hydrophobic residues are buried in the protein core and most of the hydrophilic residues prefer residing on the surface, exposed to water. Consequently, a native globular protein generally displays a compact structure, containing a hydrophobic interior. For the smaller proteins, this native structure has been proposed to be achieved via a two-state kinetic model (Fersht et al., 1994; Otzen et al., 1994). This, however, may not be the case for the larger proteins, where the initial fast step may involve collapse of the hydrophobic cores with subsequent rearrangement to yield some fully formed secondary structure elements along with tertiary structure interactions.

Alternatively, secondary structure elements may form late in the folding pathway (Otzen et al., 1994). The slower step apparently involves attaining the native conformation via the pairing of the already largely pre-folded domains or folding units. The latter steps are likely to resort to some conformational re-arrangements to optimize their packing interactions. Analysis of known protein structures indicates that a polypeptide chain with a size larger than 250 residues barely folds cooperatively into just one hydrophobic unit (Garel, 1992). This behavior is consistent with the folding free energy, in which the entropy term is in favor of the short polypeptide chain folding separately into an independent unit.

Here we provide an automated procedure to identify independent stable folding units in native protein structures. These units may be viewed as hydrophobic core nucleation sites during protein folding. Where data are available, they are consistent with intermediate folds as seen in protein kinetic studies (e.g., Freire, 1995). Hence, the procedure, and the ensemble of hydrophobic folding units it produces from protein structures, are expected to be useful

Reprint requests to: R. Nussinov, NCI-FCRF Bldg. 469, Rm. 151, Frederick, Maryland 21702; e-mail: ruthn@ncifcrf.gov.

in identifying intermediates correlated with folding kinetics. The critical role played by the hydrophobicity in a folding unit has been shown for fragments having the same fold as in the native protein (Novokhatny et al., 1984). More recently it has been illustrated in a designed 23-residues fold (Struthers et al., 1996). The hydrophobic folding units that we generate have been defined to have a size of at least 30 residues, affording tertiary interactions between partially folded elements. While these folded units possess relatively strong hydrophobic cores, their hydrophobic interactions with their surroundings, or with other units, are relatively small. They are compact, and may consist of non-contiguous segments on the polypeptide chain. Their sizes and non-contiguity afford inclusion of non-local interactions, manifested at the initiation sites. For smaller proteins, such as the chymotrypsin inhibitor 2, only one unit is generated. For larger proteins, several hydrophobic folding units are automatically obtained, and often more than one hydrophobic folding unit is generated for a domain (e.g., four units for actin), consistent with the projection that in larger proteins, formation of a domain may be preceded by a collapsed, rearranged, hydrophobic module (Otzen et al., 1994). Whereas some hydrophobic folding units are straightforwardly identified by our algorithm, others are more difficult to discern. These cases generally display appreciable hydrophobic interactions between the units and their respective neighborhoods. The implications of these cases with respect to folding kinetics are assessed and addressed in this work.

The biological function of a particular protein relies on the stability of its uniquely folded conformation, which consists of one or of a number of domains. In the latter case, each domain may fulfill its own biological function independently or in a concerted manner with its neighbors. The structural unit of the protein has been viewed and defined by a number of criteria, including structure, function, stability, genetic, thermodynamic, and evolutionary considerations (Garel, 1992). In general, domain definitions correspond to structural entities, since the function and the stability of a protein are based on its folded conformation. Gazing at three-dimensional models of protein structures illustrates that a structural domain is a compact, globular substructure with more interactions within it than with its neighborhood (Janin & Wodak, 1983). Hence, a structural domain can be described by two visual characteristics: its compactness and its extent of isolation. Although the boundary of the domain can be determined by visually inspecting protein structures, constructing a reliable automated domain-cutting procedure is not a straightforward task. In part, this is because the implicit visualization criteria of a domain cannot be directly translated into quantitative terms and, in part, due to the variety of distinct ways to encode these two properties to evaluate the quality of the domain.

A number of distinct algorithms for domain identification have been proposed in the last two decades (Crippen, 1978; Lesk & Rose, 1981; Rashin, 1981; Wodak & Janin, 1981; Segawa & Richards, 1988) and particularly in recent years (Moult & Unger, 1991; Murphy et al., 1992; Noguti et al., 1993; Zehfus, 1993; Islam et al., 1995; Siddiqui & Barton, 1995; Swindell, 1995). In general, domain or folding unit cutting algorithms consist of two steps. In the first step plausible domains are generated by either searching, cutting, or assembling. In the second step, a scoring function evaluates each candidate to yield the best cutting results. Here, we touch on some of the more recent approaches. Zehfus (1994) exhaustively generated all candidates containing at most two non-contiguous segments and evaluated these strictly in terms of compactness in the scoring function. Restricted by the exponential

increase in the number of candidates, this approach does not lend itself to searches for compact units possessing more than three segments. Holm and Sander (1994) utilized an eigenvalue analysis of a contact matrix to cut a protein. They subsequently evaluated their candidates by adhering to the criteria of structural domains, namely, maximal interactions within each unit and minimal interactions between units. Using a reciprocal averaging procedure (Hill, 1973), their approach elegantly reduced the multiple cutting-point problem to a one-dimensional search problem. Sowdhamini and Blundell (1995) employed dendrograms of contact distances between secondary structure elements to assemble folding units. These are assessed by evaluating their compactness based on contacts between units. Panchenko et al. (1996) generate "foldons," or structural modules, by analyzing the energy landscape of the protein.

Here we dismember the protein into hydrophobic folding unit(s). A unit has a compact structure, possesses a hydrophobic core, is thermodynamically stable and can fold independently. Unlike the definition of a structural domain, a hydrophobic folding unit may have many contacts with other folding units in the native structure; however, these contacts are largely hydrophilic interactions. A structural domain, or folding unit, is generally defined by maximizing the number of all interactions (hydrophobic and hydrophilic) within the unit, and minimizing all contacts between units. Hence, a structural domain may contain more than one hydrophobic folding unit. By focusing on hydrophobic interactions and seeking units displaying a relatively strong hydrophobic effect, potential stable nucleation site(s) may be obtained. For small proteins we obtain one such hydrophobic folding unit. For the larger ones, a number of units may be derived. To generate folding unit candidates, we start by following the Holm and Sander reciprocal averaging analysis (1994), reducing the multiple cutting-point problem to a one-dimensional search. However, to comply with the goal outlined above, only the hydrophobic contacts among the residues are recorded in the contact matrix, which is utilized to obtain the first non-trivial eigenvector. Since the hydrophobic clustering is implicitly expressed in terms of the amplitude of the eigenvector, we are able to cut a protein into hydrophobic clusters by selecting a cutting index value within the range of the eigenvector's amplitudes. As many trial hydrophobic folding units as twice the size of the protein (measured by their number of residues) can be generated. A candidate is assessed by a function based on four quantities: compactness, isolation, hydrophobicity, and segmentation. We use the definition proposed by Zehfus and Rose (1986) to calculate the compactness of a cut unit. The degree of isolation of a unit is based on the solvent accessible surface area (ASA) that was originally buried in the interior of the protein and is exposed to the solvent after cutting. The hydrophobicity is equivalent to the fraction of the buried non-polar ASA with respect to the total non-polar ASA. The last factor employed in the evaluation of a folding unit is the number of non-contiguous segments it contains. A local optimal score along sequential cuts, guided by the eigenvector, is considered to correspond to a hydrophobic folding unit. The cutting algorithm alternates between iterations along opposite directions of the eigenvector. It cuts and evaluates all potential folding units. A consistent optimal cut in both directions generates the final hydrophobic folding units.

Below, we describe the scoring function and the cutting algorithm. Our algorithm requires optimization of a set of parameters. To obtain these, 20 single-chain proteins have been selected as a training set. We outline the rationale for our definition of a hydrophobic folding unit and discuss the four criteria and their quanti-

fication. The results of cutting a dataset of 318 structurally non-redundant representative single-chain monomers, derived from the structural database using our amino acid sequence order-independent algorithm (Nussinov & Wolfson, 1991; Fischer et al., 1994; Tsai et al., 1996), are presented. Examples are given, and some interesting cases are described. We outline the implications of a dataset of hydrophobic folding units and its potential application to (a) studies of plausible intermediates in folding kinetics, (b) to threading, and to (c) folding versus misfolding. We have generated two datasets of hydrophobic folding units, specifically for these applications, by accounting for the different considerations involved in the relative importance of each of the quantities detailed above. In particular, in the second dataset a larger weight is attached to the extent of isolation to create folding units geared toward the latter two studies noted above.

Using our amino acid sequence order-independent structural comparison technique, we have further carried out extensive comparisons of these folding units to produce a non-redundant set of potential independently folding nuclei. While this set is useful for the applications noted above, we regard the families of the produced related clustered structures, derived from the structurally dissimilar non-redundant monomers, to be no less interesting. Each family contributes a single representative to the non-redundant set. However, the families themselves serve as a rich dataset for comparisons of structurally similar folding nuclei having either conserved residues at spatially identical sites or amino acid pair interchanges. Studies of these are expected to provide clues to the process of protein folding (Shakhnovitch et al., 1996). Our unique structural comparison methodology affords the generation of such a set, containing pieces of non-contiguous sequences from protein cores.

Generation of the hydrophobic folding units: theoretical considerations

In this section, we present the rationale and considerations leading to our hydrophobic folding units cutting procedure. This is done in two subsections. In the first, we describe the four ingredients that are utilized to evaluate a hydrophobic folding unit. Their actual implementation, in consideration of the calculation speed, is discussed in the Appendix. Next, we discuss the formula of the scoring function that we employ and how to obtain an optimal set of the associated weighting parameters. In the second subsection, we outline the mechanics of our algorithm.

The scoring function of a hydrophobic folding unit

Solvent accessible surface area, ASA

Since the scoring function relies on the ASA extensively, we first describe in some detail how the ASA is obtained. The solvent accessible surface area (ASA) of a protein is calculated by following the Lee and Richards' definition (1971), with a probe ball radius of 1.4 Å. The algorithm of Shrake and Rupley (1973) has been adopted in our implementation in a manner analogous to that described by Miller et al. (1987). The atomic radii used in the calculation of the accessible surface area have been taken from CHARMM (Brooks et al., 1983). We adopt the formal charge given in CHARMM to define an atom as polar or as non-polar, with an atom having a formal charge of less than 0.25 classed as a non-polar atom. Thus, ASA^{Non} and ASA^{Pol} refer to ASA derived from non-polar and from polar atoms, respectively.

The accessible surface area of an atom is represented by discrete surface points of a sphere. Details are given in the Appendix. For each surface point an associated surface area and a normal vector are calculated. In this work, the ASA of a residue excludes area buried by atoms from the same residue as well as by the two backbone atoms (N and O), which covalently link the residue to its neighbors. Hence, the ASA of a residue is comprised of two quantities, ASA_{Buried} and ASA_{Surf} , with the ASA_{Buried} referring to the area buried by atoms belonging to other residues, and the ASA_{Surf} referring to the exposed surface area.

Compactness, Z

We follow the Zehfus and Rose definition (1986) to calculate the coefficient of compactness of a protein, Z . Z is defined as the solvent accessible surface area of a structure, divided by its minimum possible surface area, which is the surface area of a sphere with an equal volume as the structure. Hence, Z is calculated as

$$Z = ASA_{Surf} / (36\pi VOL^2)^{1/3}. \quad (1)$$

The ASA_{Surf} is the exposed solvent accessible surface area of a protein or of a trial unit. The VOL is the corresponding volume, which is numerically calculated by an integration of all individual exposed solvent accessible surface areas. Thus,

$$VOL = \sum_i \frac{1}{3} ASA_i (N_i \cdot V_i) \quad (2)$$

where N_i is the normal of an accessible surface area ASA_i , and V_i is a vector from a reference point (an origin) to the discrete surface point. Our implementation of the volume calculation described above has been verified by a numerical cubic grid point algorithm as suggested by Zehfus (1994).

Isolation, I

After a unit is generated, the solvent accessible surface area ($ASA_{B \rightarrow E}$) that was originally buried in the interior of a protein but is exposed to the solvent after cutting is calculated. Hence, the degree of isolation is intuitively expressed as

$$I_{Non+Pol} = \frac{ASA_{B \rightarrow E}}{ASA_{unit}} \quad (3)$$

where ASA_{unit} is the solvent accessible surface area of the isolated trial unit. To further emphasize the hydrophobic effect in a hydrophobic folding unit, we exclude the polar ASA in the evaluation of the extent of the isolation of the hydrophobic folding unit. Thus, the final assessment of the isolation is

$$I = \frac{ASA_{B \rightarrow E}^{Non}}{ASA_{unit}} \quad (4)$$

where $ASA_{B \rightarrow E}^{Non}$ is the non-polar ASA that was originally buried in the interior of a protein but became exposed after cutting.

Hydrophobicity, H

The ASA of a generated unit is divided into two categories: non-polar ASA (ASA^{Non}) and polar ASA (ASA^{Pol}), according to the atom type to which the ASA belongs. The hydrophobicity is calculated as the fraction of the buried non-polar area out of the total non-polar area,

$$H = \frac{ASA_{Buried}^{Non}}{ASA_{Buried}^{Non} + ASA_{Surf}^{Non}} \quad (5)$$

where ASA_{Buried}^{Non} and ASA_{Surf}^{Non} are the buried and the exposed non-polar ASA, respectively. Also note that the hydrophobicity was *not* calculated as the fraction of the non-polar area in the total buried area,

$$H = \frac{ASA_{Buried}^{Non}}{ASA_{Buried}^{Non} + ASA_{Buried}^{Pol}} \quad (6)$$

where ASA_{Buried}^{Pol} is the buried polar ASA. Two reasons account for the choice of the former formula (Eq. 5), over the latter (Eq. 6) to express the hydrophobicity of a folding unit. In a dataset of sequence and structure, non-redundant 318 single-chain proteins derived from the PDB, the hydrophobicity calculated by Equation 6 tends to be a constant value, with little fluctuation, and therefore is not useful. Furthermore, using Equation 6, most of the generated, relatively small trial units would obtain a high hydrophobicity (H) score, even though substantial non-polar area is exposed to the solvent. On the other hand, the hydrophobicity as evaluated by Equation 5, reflects a practical situation, giving highly exposed non-polar area a lower score. Intuitively, the definition expressed in Equation 6 for the assessment of the hydrophobicity, seems reasonable, superior to Equation 5. However, from a practical standpoint, the choice of Equation 5 for expressing hydrophobicity is unreasonable. Certainly, for an extreme case where a protein has only one non-polar residue out of a 100 residues, Equation 5 will break down, assigning the trial unit either a perfect hydrophobicity score of 1.0, or 0.0, i.e., no hydrophobicity at all. However, as we are dealing with real proteins, such an extreme case is unlikely. A typical folded protein possesses a hydrophobic core, with most of the hydrophilic residues residing on the surface, exposed to the solvent.

Segmentation, S

We consider the order of the residues in the PDB file as continuous on the chain, even though some residues are not explicitly assigned in the PDB structure file. Therefore, any discontinuity that is encountered in the sequence order in the cut unit is used to calculate the number of segments, S .

Scoring function

The scoring function of a hydrophobic folding unit contains the four quantities, and has the form

$$\text{Score}^{\text{HFU}} = \frac{1}{Z-1} \cdot \frac{1}{I+W_1} \cdot H \cdot \frac{\ln(F/W_3)}{\ln(W_2/W_3)} \quad (7)$$

where W_1 , W_2 , and W_3 are weighting parameters, and F is the number of residues in the cut unit divided by its segment number, S . We have designed the scoring function to favor folding units that are compact objects (small Z); highly isolated modules (small I); highly hydrophobic domains; and have larger sizes, however, with a smaller number of segments. The balance of these four measurements relies on adjustable weighting parameters to truly reflect the quality of a folding unit.

Generation of candidates

Multiple-cutting-point algorithm

We use a scaling method known as reciprocal averaging or correspondence analysis (Hill, 1973) to generate a series of plau-

sible folding units. This method has been used by Holm and Sander (1994) for finding a binary partition with any number of cutting points on the sequence. Since this procedure has already been presented in detail, we do not repeat its description here, but refer the reader to the clear exposition of Holm and Sander (1994). The advantage of this approach is that the multiple cutting-point problem is reduced to a one-dimensional search for all reasonable trial cuts. To bring out the plausible hydrophobic core, the contact matrix we use to obtain the first non-trivial eigenvector contains only the hydrophobic contacts and the residue connectivity. In the following, we give the details of the calculation, and in each subsection, we describe the results of an example protein, Factor H (PDB code 1hfh), as a supplement for clarification.

Contact matrix

There are two distinct ways to assign hydrophobic contacts between residues. The first is based on the criterion of an atom-to-atom contact. Given a protein structure with n amino acids, the contact matrix, $C_{n \times n}$, is initially set to zero. If the distance between any two atoms of two hydrophobic residues (i, j) is less than the sum of their van der Waals radii plus 0.5 Å, the $C_{i,j}$ of the contact matrix is set to a value of 2.0. To emphasize long-range (i.e., non-local) interactions, all $C_{i,j}$ are set to 0.0 if $\text{abs}(i-j) < 4$, except those residues which are backbone connected. In this case, $C_{i,j}$ is set to 1.0. In this study, Ala, Val, Leu, Ile, Phe, Trp, Met, Pro, Gly, Tyr, and Cys are considered hydrophobic residues.

In the second approach, the non-polar ASA buried between residues is utilized in the contact matrix. We first define the $ASA_{i,j}$ to be the non-polar ASA of residue i buried by residue j . If $ASA_{i,j}$ and $ASA_{j,i}$, which reflect the hydrophobic contacts between i and j , are both larger than a preset value (25.0 Å²), $C_{i,j}$ is set to a value of $(ASA_{i,j} + ASA_{j,i})/20$. Otherwise, it is set to zero. As in the atom-to-atom contact criterion, long-range interactions are emphasized by setting all $C_{i,j}$ to 0.0 if $\text{abs}(i-j) < 4$, except backbone connected residues. For these, $C_{i,j}$ is set to 1.0. We found both matrices to be quite similar. However, in general, the second approach has a slightly better performance in cutting out hydrophobic folding units. There are two possible reasons for the superiority in the utilization of the non-polar ASA, which is buried between the residues in the contact matrix. First, hydrophobic contacts involving the hydrophobic atoms of hydrophilic residues are taken into account, for example, the aliphatic side chain of lysine. Second, whereas in a residue-by-residue approach a constant value of 2 is assigned to a hydrophobic-hydrophobic residue contact, in a surface area-based approach, the hydrophobic contacts are weighted by the extent of the surface that is in contact. We have, therefore, adopted the latter ASA criterion. Examples of both types of contact matrices of "1hfh" are given in Figure 1.

Reciprocal averaging

The solution of the reciprocal averaging procedure is shown to be equivalent to an eigenvalue-eigenvector problem of the positive semi-definite symmetric matrix,

$$D = (R^{-1/2}CR^{-1/2})(R^{-1/2}CR^{-1/2})^T \quad (8)$$

where R is a diagonal matrix of the row totals of the contact matrix C (for details, see Hill, 1973; Holm & Sander, 1994). That is, the first non-trivial eigenvector of the D matrix represents in its amplitude a new ordering, which reflects the grouping information. However, there is a trivial solution (1, 1, 1, ...) with a maximal

Fig. 1. Examples of the contact matrices for protein Factor H (PDB code 1hfh). The solid circles represent hydrophobic-hydrophobic contacts. In (a) two hydrophobic residues are defined to have a hydrophobic contact if the distance between any of their corresponding atoms is under the sum of their van der Waals radii plus 0.5 Å. In (b) the contact is based on the non-polar ASA buried between residues (see text for details). The open circles depict two residues that are backbone connected. A dot in the matrix stands for two residues whose contacts are excluded either because their sequence separation is under four residues or there is no hydrophobic-hydrophobic contact between them.

eigenvalue of 1. To avoid getting this trivial solution, one subtracts the component of the trivial solution for every trial vector in an iterative routine. Starting with a random trial vector, the eigenvalue-eigenvector solution is considered solved if either the norm of the trial vector converges with changes of less than 0.00001, or the iteration cycle reaches a maximum of 500 times. Figure 2 gives the first non-trivial eigenvector of "1hfh." A new contact matrix, whose residue order is reassigned by the amplitude of the eigenvector, is illustrated in Figure 3.

What is a cut

Since hydrophobic clustering is implicitly expressed by the amplitudes of the solved eigenvector, a cut is specified by an index value within the range of the eigenvector, dividing the protein into two parts. The first part contains those residues whose eigenvec-

Fig. 2. The first non-trivial eigenvector of the contact matrix (given in Fig. 1) that is solved by the reciprocal averaging procedure. Details are given in the text (see Holm & Sander, 1994).

Fig. 3. A new contact matrix with the residue order reassigned according to the amplitudes of the eigenvector, given in Figure 2.

tor's amplitudes are smaller than the index value, and the second part consists of those whose eigenvector's amplitudes are larger than the index value. In Figure 2, with the cut index at 0.0, the first cut part has residues 1–63, and the second part, 64–120. Obviously, a cut along the eigenvector's amplitude will produce arbitrary cutting points in the sequence. For instance, if in Figure 2 we choose the cut index at -0.09 , there are five cutting points in the sequence, and each of the two cut units consisting of three segments (for instance, 1–8; 15–28; 45–56, for the first part). With this definition, as many trial units as twice the size of the protein can be generated along the amplitude of the solved eigenvector.

(a) Reducing the number of segments: Since a cut by a one-dimensional search is not always perfect, it is necessary to resort to two additional processes to improve it. First, we join two segments that are separated by a short piece, which has not been assigned to any of the hydrophobic folding units. If the size of any

such remaining unassigned fragment of amino acids, either at the end of the chain or between the hydrophobic folding units, is less than half of the minimum length of a segment (here we use eight residues as the minimum segment length), the respective short segment is added to the trial cut unit. Hence, this procedure may fuse two cut segments into one. Second, to reduce the number of segments in the cut unit we eliminate any selected segment whose length is less than the minimum length of a segment.

(b) Combinatorial assembly: After discarding short fragments in a trial generation of a hydrophobic folding unit a risk still exists that unwanted segments may have been included. To prevent this situation, the unit is examined for all combinations of the segments. Consider a trial cut containing three segments. Six combinations of the three segments, (1), (2), (3), (1,2), (1,3), and (2,3) are generated, where 1, 2, and 3 represent the first, second, and third segments, respectively, and (1,2) implies that only the first two segments are considered to constitute the new trial cut. The combination achieving the best score in the combinatorial assembly is considered as the final cut folding unit for that cut index. If there are n segments in a unit, the total number of combinations is $\sum 2^{n-1}$ (including the one containing all segments). This procedure ensures getting the best cutting of the hydrophobic units as suggested by the hydrophobic contact matrix.

Bidirectional cutting

Along the amplitude of the eigenvector we generate a series of trial folding units as described above. Moving the cutting index from negative toward positive values and considering only the first part of a cut is termed forward cutting. Backward cutting is moving the cutting index in the opposite direction and considering only the second part of a cut as a trial cut. The cut producing a hydrophobic folding unit with a local optimal score in a series of cuts is the folding unit which is excised. The algorithm iterates alternatively along the opposite amplitudes of the eigenvector. If the number of residues left in the short linkers is less than a preset minimum (30

residues), the cutting algorithm terminates. A local optimal score is defined as follows. We temporarily keep the cut that has the highest score and count how many cuts have been executed in the follow-up trial cuts. If the score of a follow-up trial cut is higher than the registered one, we replace it with the new one and reset the follow-up count to zero. A local optimal cut is finalized if either the follow-up count is over 30 or the cutting index reaches the end. "1hfh" has been cut into two hydrophobic folding units. The scores of both the forward and backward cutting are plotted in Figure 4.

Adding unselected pieces

Since the cutting suggested by the reordered contact matrix is not perfect, some residues that are left unassigned should preferably be added to already cut units. For this reason we join the unassigned pieces to the previously cut hydrophobic folding units if it increases their score. We do this cleaning up of unassigned residues in two steps. First, we try to add, combinatorially, unassigned segments (with length >8) to all cut units, as described above. Second, only one unassigned residue is sequentially added to all existing units. We iterate through this "residue-attaching" process until there is no further increase in the score by single residue additions for any of the already cut units.

Weighting parameter optimization

Above, we have described the scoring function and the folding unit cutting algorithm. The performance of our cutting procedure has been trained and evaluated by cutting 20 selected single-chain proteins. Table 1 lists these, as well as the expected number of

Fig. 4. An example of the forward (solid line) and backward (dashed line) cutting scores of the "1hfh." This protein has been cut into two hydrophobic folding units at the cut index indicated by the arrow. Cutting at this site produces hydrophobic folding units with a local optimal score. The cutting index moves forward or backward in the plot at an interval of one residue at a time. However, if the successive trial unit is the same as in a previous cutting, its plot is not shown. The ending points of both cutting processes have been dynamically reset to the point where a folding unit has been generated. That is the reason why the backward cutting process does not pass the cutting index point set by the forward search.

Table 1. The list of 20 single-chain proteins used in scoring function optimization

PDB code	Compactness	Hydrophobicity	Score	Size	a	b	c	d
3cd4	1.727	0.81	3.46	178	2	1	2	2
1phh	1.881	0.83	0.00	394	2	3	3	3
2mcm	1.588	0.76	3.65	112	1	2	1	2
1bia	1.860	0.81	0.00	292	3	3	3	3
1sgt	1.592	0.84	4.58	223	1	2	2	2
1atnA	1.919	0.82	0.00	373	4	4	4	4
1ccr	1.571	0.76	3.77	111	1	1	1	1
1fha	1.729	0.76	3.23	172	1	1	1	1
2hhbA	1.680	0.76	3.30	141	1	1	1	1
1aak	1.746	0.80	3.22	150	1	1	1	1
1cus	1.560	0.82	4.57	197	1	1	1	1
1041A	1.707	0.80	3.47	164	2	2	2	2
3pmgA	1.947	0.84	0.00	561	3	3	3	3
5rubA	1.907	0.83	0.00	436	3	3	3	3
2aaiB	1.807	0.81	3.36	262	2	2	2	2
2abk	1.759	0.80	3.39	211	2	2	2	2
9wgaA	1.949	0.71	0.00	171	4	4	4	4
3grs	2.212	0.81	0.00	461	3	3	3	3
4tlm	1.681	0.85	0.00	316	3	2	2	3
4icd	2.041	0.81	0.00	414	3	3	3	3

See text for definitions of the compactness, hydrophobicity, and score.
a: number of folding units by visual inspection (compactness and isolation).

b: number of folding units in terms of hydrophobic clustering only.

c: number of folding units given by the authors who determined the structure.

d: number of folding units cut by our procedure.

folding units derived by visual inspection, hydrophobic clustering, and by the definition given by the authors who determined the structure. To optimize the cutting of these training proteins, adjustable parameters in the scoring function have been set as follows. First, we found that the scoring function needs to be truncated at some limit value in the four quantities mentioned above, i.e., compactness (Z), hydrophobicity (H), isolation (I), and segmentation (S). The score is set to zero if Z is greater than W_4 ; or if I is greater than W_5 ; or if H is less than W_6 ; or if the size of the cut unit (in residue units) is greater than W_2 . Second, unless a cut index already reaches the end, a local minimum is required to have a score higher than an absolute value of W_7 . Otherwise, the search for a local minimum continues. The final weighting parameters (W_i , $i =$

1,7), are listed in the legend to Table 2. The results of cutting the 20 training proteins are given in Table 2 and several selected examples are depicted in Figure 5. In general, parsing a structural domain that has a clear boundary with another domain and which contains only a single hydrophobic folding unit is relatively straightforward. A typical case is the wheat germ agglutinin (PDB code, 9wgaA). This protein contains four clear domains, each composed of a single hydrophobic folding unit. The more difficult cases are, however, those where more than one hydrophobic folding unit is contained within a domain. For instance, the hydroxybenzoate hydroxylase (PDB code, 1phh) contains two hydrophobic folding units in its second domain, which by visual inspection alone cannot be distinguished easily. The difficulty stems from the considerable

Table 2. The cutting results of the 20 proteins used in scoring function optimization

PDB Code	Z	I	H	S	Score	Residues
3cd4 (a)	1.477	0.07	0.80	1	4.282	1-97
3cd4 (b)	1.528	0.05	0.76	1	3.696	98-178
1phh (a)	1.729	0.06	0.82	4	2.556	1-69, 100-181, 269-341, 387-394
1phh (b)	1.453	0.18	0.75	1	3.164	184-264
1phh (c)	1.665	0.17	0.71	2	1.643	71-98, 342-386
2mcm (a)	1.420	0.04	0.77	3	3.664	1-35, 50-69, 94-112
2mcm (b)	1.483	0.05	0.66	2	2.446	36-49, 70-93
1bia (a)	1.478	0.13	0.70	1	2.701	244-292
1bia (b)	1.526	0.07	0.73	1	3.160	4-64
1bia (c)	1.567	0.10	0.81	1	3.782	67-243
1sgt (a)	1.675	0.10	0.78	2	2.365	1-15, 97-214
1sgt (b)	1.564	0.20	0.75	1	2.315	16-96
1atnA(a)	1.545	0.10	0.76	1	3.187	181-272
1atnA(b)	1.474	0.12	0.64	1	2.438	33-71
1atnA(c)	1.526	0.18	0.77	2	2.434	137-180, 273-340
1atnA(d)	1.563	0.13	0.79	3	2.479	1-32, 72-136, 341-373
104IA(a)	1.521	0.03	0.80	2	3.869	1-12, 69-164
104IA(b)	1.484	0.06	0.74	1	3.539	13-68
3pmgA(a)	1.562	0.09	0.82	1	3.989	1-190
3pmgA(b)	1.673	0.07	0.78	1	3.188	420-561
3pmgA(c)	1.616	0.15	0.83	1	3.218	191-419
5rubA(a)	1.470	0.13	0.75	2	2.768	349-357, 371-436
5rubA(b)	1.693	0.07	0.80	2	2.947	1-127, 276-348
5rubA(c)	1.575	0.13	0.78	1	3.141	128-275
2aaiB(a)	1.621	0.06	0.79	1	3.569	1-137
2aaiB(b)	1.505	0.07	0.79	1	4.167	139-262
2abk (a)	1.480	0.04	0.80	1	4.688	24-130
2abk (b)	1.536	0.04	0.78	2	3.423	1-20, 135-211
9wgaA(a)	1.422	0.04	0.68	1	3.772	1-41
9wgaA(b)	1.385	0.04	0.65	1	3.802	132-171
9wgaA(c)	1.412	0.07	0.70	1	3.584	42-84
9wgaA(d)	1.465	0.09	0.67	1	2.853	88-131
3grs (a)	1.686	0.14	0.76	1	2.500	143-274
3grs (b)	1.636	0.05	0.82	3	3.016	1-38, 97-141, 276-347
3grs (c)	1.749	0.12	0.74	1	2.255	348-461
4tln (a)	1.509	0.08	0.82	2	3.602	153-172, 206-316
4tln (b)	1.476	0.17	0.72	2	2.099	1-31, 53-72
4tln (c)	1.579	0.17	0.80	3	2.272	32-52, 73-152, 173-205
4icd (a)	1.740	0.10	0.79	2	2.508	1-109, 332-414
4icd (b)	1.624	0.03	0.64	1	2.311	162-195
4icd (c)	1.690	0.13	0.78	2	2.442	110-160, 197-331

The cutting results of the 20 training proteins given in Table 1. Proteins cut into only one unit are not included in the list. See text for definitions of Z (compactness), H (hydrophobicity), I (isolation), S (segmentation), and Score. The weighting parameters (W_i , $i = 1,7$), in the scoring function optimization are 0.25, 275, 0.80, 1.90, 0.25, 0.64, and 1.65, respectively.

Fig. 5. Selected cutting results from the training set of 20 single-chain proteins. The PDB codes of these chains are: 3cd4, 1phh, 2mcm, 1sgt, 1atnA, and 9wgaA. The proteins are shown in a ribbon representation with the first, second, third, and fourth cut units in red, green, yellow, and blue colors, respectively. Residues colored white are those marked as unassigned short linkers. (*Figure continues on next page.*)

Fig. 5. *Continued.*

extent of non-polar contacts between the hydrophobic folding units necessary for their association.

The algorithm we have presented here is flexible and adjustable. Adjustability implies assigning a set of optimized weighting parameters to comply with the preset cutting results of a training set.

Parsing a protein into domains by following a visual definition of a structural domain would require a different set of weighting parameters than those assigned by the criteria defined above. On the contrary, sensitivity is a disadvantage of any approach. One way to test the sensitivity of a parsing algorithm is to compare the

Table 3. A query example for the cutting results through our www page

Results of cutting 3cd4:							
Two hydrophobic folding units were found in "3cd4"							
3cd4	178	1.73	0.81	0.00	1	3.46	
Unit	Size	Compactness	Hydrophobicity	Isolation	Segment	Score	Residues List
1	97	1.48	0.80	0.07	1	4.28	1-97
2	81	1.53	0.76	0.05	1	3.70	98-178

cutting results obtained for very similar protein structures. We have chosen the lysozyme family, which has the largest number of structures in the PDB (Bernstein et al., 1977). Two families of lysozymes has been clustered in the database. The first family is comprised of 191 members having 164 residues each, and the second consists of 76 members each containing 129 residues. All 191 lysozymes in the first family have been cut into two similar hydrophobic folding units, while all 76 lysozymes in the second family have been cut into a single folding unit, showing the algorithm to produce reliable and consistent results.

Since parsing methods differ in their views as to what constitutes a folding unit, or a domain, it is difficult to carry out a comparison between them. For example, the crystallographers (Kabsch et al., 1990) who have determined the structure of actin (PDB code 1atnA) have described it as a two-domain protein, with each domain further divided into two subdomains (Fig. 5). With our algorithm, actin has been cut into four hydrophobic folding units. Nevertheless, we believe our definition of a folding unit and the rationale behind it is reasonable.

Some attributes of the hydrophobic folding units

Since our cutting algorithm is fast, we have been able to apply it to all protein monomer structures in the Protein Data Bank (Bernstein et al., 1997). The hydrophobic folding units that have been generated for every chain have been placed on the world wide web.⁴ Table 3 illustrates an example of a query for an immunoglobulin, CD4 (PDB code 3cd4) through our www page. For each chain the table enumerates its hydrophobic folding unit(s) (one or more), and the corresponding calculated compactness, hydrophobicity, isolation, and segmentation values. The residue listing contained in each is given as well. Here, we present the results obtained for the representative dataset of 318 structurally dissimilar single-chain proteins. To constitute a concise list, Table 4 notes only the protein size, the number of unassigned residues, the number of generated hydrophobic folding units, and the list of residues in each unit. Note that in some proteins there is a large number of residues that are left unassigned. These residues cannot constitute an independent folding unit, as they do not surpass the threshold set for constituting a unit. Joining them to existing unit(s) did not improve the respective scores of the latter. These amino acids may be stabilized by their interactions with the hydrophobic folding units. Figure 6 plots the distribution of the compactness of the generated hydrophobic units as a function of their size, given in Table 4. The hydrophobic folding units range in size between 30 to 275 residues, while their compactness values spread between 1.35

and 1.90. The distribution of their hydrophobicities with respect to size is depicted in Figure 7. The hydrophobicities range from 0.64 to 0.87 (by their definition, both the compactness and the hydrophobicity are unitless). The extent of the compactness and of the hydrophobicity is independent of the number of units excised from the chains. However, while, as might be expected, the compactness generally increases with the unit size, albeit with a fairly large spread (Fig. 6), the behavior observed with respect to the hydrophobicity in Figure 7 indicates that above a unit size of 125-150 residues, the extent of the hydrophobicity is constant. This behavior is consistent with the hydrophobic effect being the dominant driving force in protein folding. A histogram of the number of segments in each unit is plotted in Figure 8. Hydrophobic folding units clearly favor a small number of segments. The extent of isolation, with respect to the size of the hydrophobic folding unit, displays an even spread (not shown).

We have also placed at the www address noted in the previous footnote a dataset of non-redundant hydrophobic folding units, generated from the dataset of the dissimilar monomers. A listing of the family members associated with each of the hydrophobic folding unit representatives is also provided. Using the Geometric Hashing sequence order independent structural comparison technique, one can superimpose the family members to derive the spatially conserved and the interchanged amino acid pairs. This work is currently in progress. A related analysis for the dataset of monomers has already been published (Naor et al., 1996).

Discussion

To date, there is ample evidence that the driving force for protein folding is the hydrophobic effect (Dill, 1990). It is also well established that in large enough proteins, there are independent substantially hydrophobic initiation sites, which are formed both by contiguous amino acids, as well as by non-local interactions between residue-residue pairs (e.g., Fersht, 1994, 1995; Otzen et al., 1994). The initial collapsed nuclei may rearrange to optimize their packing and van der Waals interactions, resulting in hydrophobic folding units. The structures of such hydrophobic folding entities are likely to be similar to those observed in native proteins. To achieve the full native conformation, these folding units, if there is more than one, may seek a better arrangement to further stabilize the protein structure.

According to the protein folding model described above, a hydrophobic folding unit is defined as a compact unit containing a hydrophobic core. While there are relatively few contacts between structural domains, that is not the case for the hydrophobic folding units. Here, substantial contacts between hydrophobic units (as, e.g., in the case of one domain with two or more folding units)

⁴Can be accessed via the world wide web, at "http://www-lmmb.ncicfcrf.gov/~tsai/hfu.html".

Table 4. The cutting results of 318 representative monomer proteins

laak	150	0	1	(1-150)
laba	87	0	1	(1-87)
laco	753	105	3	(204-224;310-512) (540-746) (1-69;141-203;225-309)
lacr	77	0	1	(1-77)
ladn	92	19	1	(1-73)
lads	315	17	2	(1-41;184-298) (42-183)
laizA	129	0	1	(1-129)
laikA	449	73	3	(163-197;210-234) (50-76;265-292;325-347;370-414) (30-49;78-162;293-324;348-369;416-449)
lamp	291	15	1	(1-36;41-101;113-291)
laozA	552	0	3	(126-331) (332-552) (1-125)
lapmE	341	0	2	(1-24;84-95;116-308) (25-83;96-115;309-341)
laps	98	5	1	(6-98)
larb	263	0	1	(1-263)
larp	336	0	2	(1-145;266-294) (146-265;295-336)
lart	396	68	3	(52-120;255-299) (317-396) (121-254)
last	200	16	2	(10-98) (99-193)
lata	62	30	1	(19-27;40-62)
lavhA	318	26	3	(90-245) (9-71) (246-318)
lbhA	131	10	2	(18-51;91-121) (1-13;55-88;123-131)
lbbo	56	8	1	(1-48)
lbbt3	220	43	1	(44-220)
lbds	43	0	1	(1-43)
lbt	107	0	2	(10-47;75-97) (1-9;48-74;98-107)
lbgc	158	0	1	(1-158)
lbgh	85	14	1	(1-13;28-85)
lbja	292	5	3	(244-292) (4-64) (67-243)
lbllE	481	29	4	(200-226;298-345;398-439) (1-67) (159-199;227-297;346-397;440-453;463-481) (85-155)
lbmdA	327	0	3	(1-87) (163-181;193-222) (88-162;182-192;223-327)
lbmv1	185	0	1	(1-185)
lbmv2	374	17	2	(182-374) (18-181)
lbnh	456	48	3	(377-456) (1-149) (198-376)
lbpe	284	96	2	(215-284) (97-214)
lbrd	170	0	1	(1-170)
lbrnL	108	0	1	(1-108)
lbrsF	89	0	1	(1-89)
lbus	57	5	1	(1-43;49-57)
lbvh	157	0	1	(1-157)
lbw3	125	0	1	(1-125)
lbyb	490	0	2	(1-102;152-174;277-288;372-460;479-490) (103-151;175-276;289-371;461-478)
lc2rA	116	3	2	(42-96) (1-41;100-116)
lc5a	65	6	1	(1-59)
lcAuB	184	50	1	(1-134)
lcbh	36	0	1	(1-36)
lcbn	46	0	1	(1-46)
lcde	209	54	2	(1-104) (108-135;151-173)
lcdg	686	74	5	(407-498) (586-686) (499-582) (139-212) (1-135;281-406)
lcdlA	142	10	2	(1-69) (80-142)
lchc	68	11	1	(1-57)
lchmA	401	0	2	(156-401) (1-155)
lcmbA	104	0	1	(1-104)
lcolA	197	0	1	(1-197)
lcoy	501	180	2	(1-40;228-314;445-472;488-501) (154-214;320-378;406-437)
lpcB	172	0	2	(52-87) (1-51;88-172)
lcpn	208	0	1	(1-208)
lcseE	274	0	1	(1-274)
lcseI	63	0	1	(1-63)
lcsh	435	91	2	(275-391) (62-274;392-405)
lctf	68	0	1	(1-68)
lctm	250	0	2	(1-169;230-250) (170-229)
lcus	197	0	1	(1-197)
lcyhA	164	0	1	(1-164)
ld66A	57	23	1	(1-34)
lddt	523	4	4	(1-13;89-136) (368-523) (14-88;137-185) (190-367)
ldcc	39	0	1	(1-39)

(continued)

Table 4. Continued

ldlhB	188	9	2	(91–188) (1–81)
ldsbA	188	4	2	(1–65;138–188) (68–135)
lede	310	9	2	(116–162;229–310) (1–109;165–227)
left	405	9	4	(27–46;171–204) (215–312) (9–26;47–170;205–213) (313–405)
legf	53	21	1	(1–32)
leipA	169	0	1	(1–169)
lepg	53	5	1	(6–53)
lesl	157	37	1	(1–120)
letu	177	10	1	(6–155;161–177)
lf3g	150	0	1	(1–150)
lfbaA	360	116	2	(105–297) (34–62;305–326)
lfc2C	43	0	1	(1–43)
lfc2D	206	4	2	(105–206) (1–100)
lfcba	494	129	2	(1–97) (98–151;179–360;411–442)
lfcbb	400	246	2	(94–179) (1–46;321–342)
lfd2	106	0	1	(1–106)
lfha	172	0	1	(1–172)
lfiaA	79	39	1	(40–79)
lfkb	107	0	1	(1–107)
lfrd	98	0	1	(1–98)
lgal	581	170	2	(1–55;224–321;435–446;529–549;565–581) (179–213;322–434;447–506)
lgatA	60	8	1	(1–52)
lgl1O	334	0	3	(175–208;229–236) (1–150;315–334) (151–174;209–228;237–314)
lgdhA	320	15	3	(196–281) (1–99;296–320) (101–195)
lgky	186	1	3	(122–152) (33–81) (1–32;83–121;153–186)
lglaG	489	35	3	(1–235) (302–334;354–373) (236–301;374–473)
lgluA	81	6	1	(7–81)
lgnc	175	34	1	(25–165)
lgof	639	16	4	(264–435) (6–150) (542–577;583–639) (157–263;436–541)
lgp1A	183	0	1	(1–183)
lgpb	823	230	4	(1–44;86–107) (703–763) (272–464) (465–702;764–798)
lgph1	465	100	2	(1–228) (261–375;395–416)
lhbq	176	0	1	(1–176)
lhcrA	52	17	1	(5–39)
lhev	43	0	1	(1–43)
lhfh	120	1	2	(1–64) (66–120)
lhgeA	328	119	1	(59–267)
lhgeB	175	56	2	(119–175) (1–49;106–118)
lhip	85	0	1	(1–85)
lhme	77	30	1	(11–57)
lhmy	327	0	2	(1–186;306–327) (187–305)
lhoe	74	4	1	(5–74)
lhplA	449	61	2	(1–276) (338–449)
lhasA	276	1	3	(111–162) (182–276) (1–110;163–180)
lhasbB	99	0	1	(1–99)
lhstA	74	0	1	(1–74)
lhuw	166	0	1	(1–166)
lhyp	75	13	1	(1–62)
liag	201	0	1	(1–201)
lifc	131	0	1	(1–131)
liffhH	219	1	2	(1–119) (121–219)
ligeA	322	3	3	(210–322) (1–107) (111–209)
lkst	68	0	1	(1–68)
ll01	164	0	2	(1–12;67–164) (13–66)
llab	80	0	1	(1–80)
llba	146	0	1	(1–146)
llccA	51	0	1	(1–51)
llcf	691	38	5	(438–459;527–591) (92–250) (460–526) (1–91;251–310) (341–437;592–683)
llct	324	9	2	(89–247) (1–87;249–317)
llis	131	0	1	(1–131)
llla	600	122	4	(1–30;46–98) (189–202;354–389;416–490;516–544;564–600) (99–130;393–413) (203–353)
llmb4	92	13	1	(14–92)
llst	238	2	2	(1–89;194–238) (91–192)
llz1	130	0	1	(1–130)

(continued)

Table 4. Continued

lmat	263	0	1	(1-263)
lmbd	153	0	2	(1-28;66-153) (29-65)
lmcoH	428	5	4	(326-428) (221-321) (1-117) (119-220)
lmcoL	216	0	2	(1-111) (112-216)
lmcpL	220	0	2	(1-113) (114-220)
lmfa	229	1	2	(1-112) (114-229)
lminA	468	16	4	(49-194) (1-33;383-405) (195-315) (317-382;406-468)
lminB	522	103	3	(84-92;113-188) (357-469) (68-83;93-112;189-351;470-491)
lmjc	69	5	1	(1-55;61-69)
lmp	357	48	2	(1-182) (208-334)
lmycC	462	250	3	(63-86;141-172) (318-372) (87-140;229-275)
lnar	289	0	2	(188-289) (1-187)
lnbaA	253	0	1	(1-253)
lnbtA	66	0	1	(1-66)
lndc	150	0	1	(1-150)
lnipB	287	0	2	(1-145;214-251) (146-213;252-287)
lnxb	62	0	1	(1-62)
loctC	131	4	2	(1-71) (76-131)
lofv	169	0	1	(1-169)
losa	148	26	2	(15-68) (81-148)
lovaA	385	43	2	(32-88;100-179;290-330) (180-288;331-385)
lovb	159	0	1	(1-159)
lpba	81	7	1	(8-81)
lpda	296	0	3	(89-189) (1-88;190-210) (211-296)
lpdc	45	0	1	(1-45)
lpdgA	87	14	2	(31-50;75-87) (14-29;51-74)
lpfkA	320	14	2	(1-143;253-306) (144-252)
lpgx	70	7	1	(1-63)
lphh	394	8	3	(1-69;100-181;269-341;387-394) (184-264) (71-98;342-386)
lphp	394	1	2	(177-383) (1-176;385-394)
lpi2	61	0	1	(1-61)
lpil	452	7	5	(1-35;115-160) (256-358) (161-216) (359-445) (36-114;217-255)
lpkp	145	0	1	(1-145)
lpmy	123	0	1	(1-123)
lpoc	134	0	2	(1-32) (33-134)
lpod	124	0	2	(1-14;46-95) (15-45;96-124)
lpom	51	5	1	(6-51)
lpoxA	585	53	3	(1-181) (186-346) (352-541)
lppt	36	0	1	(1-36)
lprcC	332	150	1	(29-149;272-332)
lprcH	258	162	1	(133-228)
lprcM	323	274	1	(216-264)
lpr	173	2	2	(1-87) (90-173)
lpspA	106	30	2	(11-49) (62-98)
lptf	87	0	1	(1-87)
lptsA	119	17	1	(1-94;112-119)
lpxtB	348	125	1	(1-132;176-193;207-236;306-348)
lpyaB	228	41	3	(75-96;117-135;161-174) (7-16;44-70;179-228) (22-41;136-160)
lpyp	280	26	1	(1-254)
lr091	273	155	1	(60-90;110-172;206-229)
lraaA	310	0	2	(1-140;292-310) (141-291)
lrabD	153	8	2	(107-153) (1-98)
lrcb	129	0	1	(1-129)
lrec	185	2	2	(85-185) (1-82)
lrhd	293	8	2	(6-157) (161-293)
lribA	340	12	3	(162-192) (47-88;113-153;203-239;278-314;324-340) (1-44;90-111;193-202;240-277;315-323)
lrip	81	16	1	(1-22;39-81)
lrpa	342	0	2	(1-15;38-138;211-342) (16-37;139-210)
lrro	108	0	1	(1-108)
lrte	268	1	2	(1-35;46-210) (37-45;211-268)
lrveA	244	1	3	(205-244) (18-34;141-162) (1-17;35-139;163-204)
lsacA	204	0	1	(1-204)
lsbp	309	0	2	(96-219;276-309) (1-95;220-275)
lscc	482	275	2	(1-86) (170-290)

(continued)

Table 4. Continued

lscmC	149	2	2	(1-75) (78-149)
lscs	237	0	1	(1-237)
lsesA	421	26	4	(129-168;391-421) (47-80) (169-390) (1-46;81-102)
lshl	48	0	1	(1-48)
lshaA	103	0	1	(1-103)
lsis	35	0	1	(1-35)
lsltA	133	0	1	(1-133)
lsrdA	154	0	1	(1-154)
lsto	208	0	1	(1-208)
ltahA	318	0	2	(1-114;168-211;266-318) (115-167;212-265)
ltbpA	180	8	2	(99-180) (9-98)
ltca	317	0	2	(49-61;180-278) (1-48;62-179;279-317)
lthg	544	219	3	(1-63;97-166;185-251) (359-426) (167-184;266-304)
ltie	166	0	1	(1-166)
ltml	286	0	3	(75-96;118-139) (13-35;262-286) (1-12;36-74;97-117;140-261)
ltndA	323	29	3	(243-273) (44-133) (1-30;150-242;274-323)
ltnfA	152	0	1	(1-152)
ltnrR	139	24	2	(1-58) (83-139)
ltop	162	8	2	(1-89) (98-162)
ltpfA	250	4	2	(5-124;233-250) (125-232)
ltplA	426	21	2	(56-307) (14-51;312-426)
ltrb	316	3	2	(1-115;245-316) (118-243)
ltrkA	678	15	3	(1-123;145-160;176-310) (540-678) (124-144;311-539)
lttaA	127	0	1	(1-127)
lula	289	51	2	(1-112) (113-146;171-223;237-275)
lutg	70	13	1	(1-57)
lvil	126	9	1	(1-117)
lvmoA	163	0	1	(1-163)
lvsgA	362	0	3	(1-12;108-144;173-226) (40-77;263-362) (13-39;78-107;145-172;227-262)
lwhA	256	3	2	(1-185) (189-256)
lwhB	153	51	1	(52-153)
lwsyA	248	36	2	(19-127;216-248) (146-215)
lwsyB	385	78	3	(83-179) (51-71;319-385) (197-318)
lxib	388	158	2	(51-215) (9-50;282-304)
lysc	421	25	3	(1-174;270-303) (180-264;336-346) (317-335;349-421)
lzaaC	85	17	2	(1-31) (49-85)
2aaa	476	100	3	(121-187) (364-476) (1-120;188-250;340-352)
2aaiB	262	1	2	(1-137) (139-262)
2abxA	74	0	1	(1-74)
2act	218	9	2	(1-13;116-214) (17-113)
2bat	388	9	2	(96-235) (10-95;236-388)
2bb2	181	6	2	(93-181) (1-86)
2bbkH	355	109	2	(186-355) (110-185)
2bbkL	125	6	1	(7-125)
2bbqA	264	0	1	(1-264)
2bbvC	321	10	2	(155-176;220-232) (1-153;177-218;234-259;267-321)
2bopA	85	0	1	(1-85)
2bpa1	426	144	3	(73-125) (297-355) (1-72;235-292;387-426)
2bpa2	175	25	1	(16-165)
2bpk	131	30	1	(1-40;60-120)
2cba	258	24	1	(25-258)
2cdv	107	45	1	(46-107)
2chsA	114	0	1	(1-114)
2ctc	307	39	1	(1-131;171-307)
2drpD	65	31	1	(1-34)
2ech	49	0	1	(1-49)
2gfl	70	0	1	(1-70)
2glsA	468	172	3	(1-103;225-236) (130-224;237-267) (273-302;332-356)
2gstA	217	30	2	(87-190) (1-83)
2hpdA	457	38	3	(1-68;330-369) (71-125;146-268) (126-145;269-329;406-457)
2ilb	153	0	1	(1-153)
2lbp	346	1	2	(1-120;252-329) (121-250;330-346)
2mcm	112	0	2	(1-35;50-69;94-112) (36-49;70-93)
2mhr	118	0	1	(1-118)

(continued)

Table 4. Continued

2mtaC	147	0	1	(1-147)
2ohxA	374	0	2	(1-178;318-374) (179-317)
2pf2	135	19	2	(67-130) (1-52)
2pkaA	80	10	1	(11-80)
2pkaB	152	54	1	(37-134)
2plv2	268	27	1	(28-268)
2polA	366	4	3	(1-117) (249-366) (118-244)
2por	301	78	2	(28-84;104-187) (6-26;87-100;237-248;263-283;288-301)
2reb	303	22	3	(78-191) (245-303) (21-77;194-244)
2rn2	155	0	2	(66-126) (1-65;127-155)
2rsIB	120	0	2	(64-120) (1-63)
2rspA	115	0	1	(1-115)
2scpA	174	0	1	(1-174)
2sicI	107	0	1	(1-107)
2sil	381	4	3	(118-143;158-223) (224-381) (5-117;144-157)
2sn3	65	5	1	(1-60)
2sns	141	0	1	(1-141)
2snv	151	0	2	(1-66) (67-151)
2stv	184	16	1	(16-183)
2tbvC	321	0	2	(1-206) (207-321)
2tgf	50	5	1	(1-45)
2tgi	112	0	1	(1-112)
2tmdA	729	201	3	(57-262) (494-640) (383-493;641-704)
2tmnE	316	0	2	(152-173;204-316) (1-151;174-203)
2tmvP	154	2	2	(1-18;54-70;138-154) (19-52;71-136)
2tsI	317	0	2	(244-317) (1-243)
2wrpR	104	47	1	(48-104)
2yhx	457	36	3	(60-185) (285-359) (18-59;186-256;271-281;362-457)
3aahA	571	45	2	(167-178;190-389;402-457) (1-166;179-189;390-401;468-527;563-571)
3b5c	85	0	1	(1-85)
3bcl	344	211	2	(132-205) (286-344)
3cd4	178	0	2	(1-97) (98-178)
3chy	128	0	1	(1-128)
3cla	213	0	1	(1-213)
3dni	258	8	2	(89-230) (1-86;231-252)
3ecaA	326	10	2	(1-201) (212-326)
3gapA	208	5	2	(1-131) (137-208)
3grs	461	60	3	(143-274) (1-38;97-141;276-347) (348-461)
3hhrB	195	0	2	(91-195) (1-90)
3mdsA	203	12	2	(6-34;81-203) (39-77)
3pgm	230	0	1	(1-230)
3rubS	123	30	1	(11-46;67-123)
4cpaI	37	5	1	(6-37)
4enI	436	0	3	(156-167;216-225;249-344;357-371) (1-138;345-356;372-386) (139-155;168-215;226-248;387-436)
4gcr	174	8	2	(84-169) (1-80)
4icb	76	0	1	(1-76)
4mt2	61	30	1	(1-31)
4ptp	223	0	2	(1-8;106-212) (9-105;213-223)
4rxn	54	0	1	(1-54)
4sgbI	51	0	1	(1-51)
4tms	316	0	3	(1-54;215-234;248-282;301-316) (87-139) (55-86;140-214;235-247;283-300)
5fbpB	314	0	2	(1-178) (179-314)
5pti	58	0	1	(1-58)
5rubA	436	13	3	(349-357;371-436) (1-127;276-348) (128-275)
6insE	50	6	1	(7-50)
7icd	414	42	4	(35-61;312-414) (161-196) (1-34;62-102) (145-160;197-311)
7rsa	124	0	1	(1-124)
8abp	305	8	2	(109-253;285-297) (1-108;254-284)
8dfr	186	0	1	(1-186)
8fabA	206	2	2	(1-103) (106-206)
9rnt	104	0	1	(1-104)
9wgaA	171	3	4	(1-41) (132-171) (42-84) (88-131)

In each line, the first column is the PDB code and the chain identifier. The second column details the protein size (in residues). The third column is the number of unassigned residues. The fourth column indicates the number of hydrophobic folding units that have been determined for the chain. The last column lists the residues (in parentheses) for each cut unit. Note that the specified residue list of a folding unit is given by reference to the residue order of a chain in the PDB file.

Fig. 6. A plot of the distribution of compactness as a function of unit size, for 552 hydrophobic folding units, which are cut from a representative dataset of 318 monomer proteins. The open and solid circles represent, respectively, only one unit and at least two units being cut from a chain.

cannot be excluded. However, these contacts are largely the outcome of interactions between hydrophilic groups.

The intermediate stage in protein folding can thus be viewed as an assembly of stable subdomain entities (Wu et al., 1994). Such subdomains have already been defined as portions of the proteins that may be larger than individual secondary structure elements but smaller than domains (Rose, 1979; Wu et al., 1994). Moreover, it has been proposed that the interactions between these subdomain entities can be likened to intramolecular recognition in much the same way as intermolecular recognition and association. It may be reasonable to assume that, on average, the relative fraction of the surface area buried between the interacting subdomains is larger than that buried between interacting subunits in oligomeric proteins. In particular, however, a main determinant of protein association is the non-polar buried surface area. Hence, it is also reasonable to assume that for the case where the two subunits fold separately and associate as relatively rigid bodies, the hydrophobic interactions between the associating subunits are fewer than in the case of the associating hydrophobic folding units. On the other hand, the extent of the hydrophilic interactions between the hydrophobic folding units is smaller than that observed at protein-protein interfaces. The considerations and implications of the association of two (or more) hydrophobic folding units are in many respects those observed between associating domains. Faulty association of hydrophobic folding units would result in misfolding, leading to protein aggregation. Alternate binding modes of the hydrophobic folding units may follow a mechanism similar to that recently proposed for domain swapping (Bennett et al., 1995).

In our procedure for generating the hydrophobic folding units we allow nonconsecutive fragments of the chain to be included in the same unit. This approach is borne out both by visual inspection and by hydrophobicity considerations. In addition, however, recent results indicate that chain connectivity is not necessarily fold determining. Proteolytic dissection of the native trp repressor and of horse heart cytochrome *c*, yielded cleaved fragments, capable of undergoing spontaneous noncovalent association, to form the native subdomain conformation (Wu et al., 1994). These results are consistent with relaxing the definition of a subdomain to include chain segments that are not covalently connected.

Fig. 7. A plot of the distribution of hydrophobicity as a function of unit size, for 552 hydrophobic folding units, which are cut from a representative dataset of 318 monomer proteins. The open and solid circles represent, respectively, only one unit and at least two units being cut from a chain.

Fig. 8. A histogram of the number of segments in the 552 hydrophobic folding units derived from a representative dataset of 318 monomer proteins.

Nevertheless, chain connectivity cannot be entirely disregarded in the definition of a subdomain. First, based on the probability that the protein chain would fold back upon itself to form a unit built from several interacting nonconsecutive segments, it is reasonable to assume that a hydrophobic folding unit possessing a smaller number of contiguous fragments would be more favorable than one with a larger number of nonconsecutive segments. Chain connectivity counteracts entropy during folding, by increasing the effective local concentration of the amino acids, with the interactions between the latter determining the adopted fold.

Our definition of a hydrophobic folding unit includes four ingredients: compactness, extent of isolation, hydrophobicity, and segmentation. In what follows, we illustrate why it is necessary to include all four measurements in parsing a protein into such units. Examples demonstrating the role played by these ingredients are provided for the 20 training proteins (Fig. 5 and Table 2). First, had we used only the criterion of hydrophobic clustering, a molecule such as the immunoglobulin CD4 (PDB code 3cd4) will not be cut into its two units. As illustrated in Figure 9, depicting only the side

Fig. 9. A ribbon representation of the immunoglobulin CD4 (PDB code 3cd4). Only the buried hydrophobic residues are shown in a ball and stick model to indicate the continuation of the hydrophobic residues within the protein.

chains of the buried hydrophobic residues, this is owing to the continuation of these from one unit to its adjacent one. While no clear boundary is detected between two separate hydrophobic clusters, visual inspection via backbone trace only, still distinguishes two domains or folding units. Second, visually the two hydrophobic folding units of trypsin (PDB code 1sgt) are not identified as separate entities, since there is an appreciable amount of contact between them (Fig. 5). This implies that using a visual definition of a structural domain, without including a measure of hydrophobicity, only one unit would be assigned for trypsin. Third, the balance between all four measurements, namely, compactness, isolation, hydrophobicity, and segmentation enabled the detection of two hydrophobic folding units in macromomycin (PDB code 2mcm). Fourth, for the case of barnase, the three clearly discerned hydrophobic clusters correspond to a single hydrophobic folding unit. This is consistent with the experimental evidence showing barnase to follow a two-state folding model (Griko et al., 1994). Incorrectly parsing barnase into three independent units would have resulted in too much exposed non-polar surface area at their interface, rendering these subdomain-nucleation centers unstable entities. Our algorithm considers barnase a single entity since parsing it into two units, the scores of hydrophobicity and of isolation are highly unfavorable. Lastly, employing solely the criteria of compactness and hydrophobicity, a 4-helix bundle would have been spliced across the helices into two units, with the cut roughly perpendicular to the axes of the helices. Inclusion of a segmentation criterion in the definition of the hydrophobic folding unit circumvents such an undesirable excision.

The concept underlying our definition of a hydrophobic folding unit requires that it be both stable and independent. On the one hand, such a concept corresponds to an “all-or-none” description of a protein fold. If any part of the unit is not included in the fold, the unit is unstable, and as such would not constitute an independent entity. Cases where domain swapping have been demonstrated provide some illuminating examples (Bennett et al., 1995).

There, a protein fold is maintained regardless of whether the associating domains are from the same chain or from different ones or whether the domains are oriented in the native or in non-native orientation with respect to each other. On the other hand, at the same time, the algorithm enables parsing a protein in the middle of a secondary structure element (e.g., a long helix linking two domains in cd4) or, between interacting secondary structure elements forming a fold, as in the middle of a β -sheet (e.g., 1phh). These considerations are consistent with the hydrophobic folding units being nucleation sites during protein folding. Nucleation sites have been proposed to contain weak tertiary interactions between partially formed secondary structures from more than one segment of the protein. These sites are likely to be both stable and independent folding units.

A comparison between protein–protein interfaces and the interfaces between independent folding units can be particularly instructive. The association between these units during protein folding is not unlike the association of protein monomers. Since in the latter case different chains are involved, one may obtain insight from protein–protein associations to better define hydrophobic folding units and locate potential nucleation sites. Examination of the associations of the hydrophobic folding units illustrates some of the similarities with, and differences from, the protein–protein interfaces we have assembled previously (Tsai et al., 1996, 1997). The architectures across the interfaces between the hydrophobic folding units resemble those seen at the protein–protein (subunit) interfaces. For example, β -sheets with strands contributed by two adjacent folding units are frequently observed. A related interesting example concerns the α/β type fold. Examination of our dataset of protein–protein interfaces (Tsai et al., 1996) has shown that across subunit boundaries the α/β type architectures are formed where one helix, along with part of a β -sheet is donated by one subunit, while a second helix with another part of a β -sheet is contributed by the neighboring subunit. In none of the protein–protein interfaces we have inspected one subunit donates a sheet, whereas its

counterpart contributes the helices. Such a situation has not been observed in the interfaces between our hydrophobic folding units either. In particular, our cutting algorithm does not excise a helix from its neighboring sheet, as the hydrophobic interactions between packed α/β folds are too favorable to permit such cutting. By inference, since protein subunits fold separately with subsequent assembly into oligomeric structures, this may lend support to an alternate folding pathway where, in general, distinct secondary structure elements do not form independent initiation sites. Rather, the similarity in the contributions of the "part-architectures" by the modules, whether monomers or hydrophobic folding units, is consistent with a pathway where partially formed secondary structures, interacting with their neighbors, coalesce to form a domain. On the other hand, the highly hydrophobic 4-helix bundles observed at the interior of proteins are not cut longitudinally. That is unlike the situation at protein-protein subunit interfaces, where typically two helices are contributed by each subunit. Studies of protein-protein interfaces, where the two subunits have folded separately and associated as relatively rigid units, provide clues to what constitutes a nucleation site. However, in cases where (unfolded) subunits fold cooperatively, the interfaces may reflect architectures and interactions observed in the interior of independently folding hydrophobic units.

Conclusions

We have used the concept of hydrophobic clustering combined with a visual definition of a domain (i.e., compactness and isolation) to design an automated, fast algorithm for parsing a protein into hydrophobic folding units. The hydrophobic clustering has been implicitly embedded in a multiple cutting-point procedure. The success of the cutting algorithm relies on the inclusion of suitable measurements for evaluating the quality of a trial cut unit and on the balance of these in the scoring function.

A description of a protein structure in terms of hydrophobic folding units has several potential implications: first, identification of hydrophobic folding units is projected to aid in locating nucleation sites during the process of protein folding. Second, a hydrophobic folding unit is a reasonable candidate for an intermediate fold as seen in studies of protein kinetics or of the molten globule state (Privalov, 1996). Third, a folding unit may constitute a unique structural motif having a particular associated protein function. Fourth, we can simplify protein architecture according to the classification of its hydrophobic folding units. Hydrophobic folding units can be unambiguously classified into three structural classes, all- α , all- β , and α - β . Protein structure can consequently be classified into four classes as all- α , all- β , $\alpha + \beta$, and α/β classes. The $\alpha + \beta$ class corresponds to a protein composed of at least one all- α and one all- β hydrophobic folding units. The α/β class consists of α - β hydrophobic folding unit(s) containing both helices and β -strands. The parsed results further indicate that a polypeptide chain folds into a compact hydrophobic unit with as few as 40 residues. A compact protein can be composed of two (or more) individual hydrophobic units.

Our sequence order independent structural comparison technique, has afforded compilation of a unique representative dataset of structurally non-redundant hydrophobic folding units. Each folding unit represents a family of clustered structurally similar folding entities. The families are particularly interesting, as they are generated from a dataset of sequence and structure dissimilar monomers. Hence, they provide a rich set of data for studies of the

mechanism of protein folding. On the other hand, the dataset of the structurally dissimilar representatives is expected to be useful for fold recognition via threading. Furthermore, a hydrophobic folding unit may be a good starting template in protein design. The fact that a similar pattern of association is observed between the hydrophobic folding units and between monomers validates our definition of hydrophobic folding units. It is also expected to prove valuable for comprehending protein-protein association.

Acknowledgments

We thank Dr. Zehfus for providing the ASA look-up table source program. We thank Drs. Robert Jernigan, Shuo Lin, Jacob Maizel, Haim Wolfson, and Dong Xu for helpful discussions, encouragement, and interest. We thank the personnel at the Frederick Cancer Research and Development Center for their assistance. The research of R. Nussinov has been sponsored by the National Cancer Institute, DHHS, under Contract No. 1-CO-74102 with SAIC, and in Israel, by a grant from the Israel Science Foundation administered by the Israel Academy of Sciences, and by a grant from the Rekanati foundation. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

References

- Bennett MJ, Schlunegger MP, Eisenberg D. 1995. 3D domain swapping: A mechanism for oligomer assembly. *Protein Sci* 4:2455-2468.
- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein databank: A computer-based archival file for macromolecular structures. *J Mol Biol* 112:535-542.
- Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J Comp Chem* 4:187-217.
- Crippen GM. 1978. The three structural organization of proteins. *J Mol Biol* 126:315-332.
- Dill KA. 1990. Dominant forces in protein folding. *Biochemistry* 31:7134-7155.
- Fersht AR. 1994. Characterizing transition states in protein folding: An essential step in the puzzle. *Curr Opin Struct Biol* 5:79-84.
- Fersht AR. 1995. Optimization of rates of protein folding: The nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci USA* 92:10869-10873.
- Fischer D, Wolfson H, Lin SL, Nussinov R. 1994. 3-D, sequence-order independent structural comparison of trypsin against the crystallographic database reveals active site similarities to subtilisin-like and sulfhydryl proteases: Potential implications. *Protein Sci* 3:769-786.
- Freire E. 1995. Thermodynamics of partly folded intermediates in proteins. *Annu Rev Biophys Biomol Struct* 24:141-165.
- Garel JR. 1992. Folding of large proteins: Multidomain and multisubunit proteins. In: Creighton TE, ed. *Protein Folding*. New York: W. H. Freeman and Company. pp. 405-454.
- Griko YV, Makhadze GI, Privalov PL, Hartley RW. 1994. Thermodynamics of barnase unfolding. *Protein Sci* 3:669-676.
- Hill MO. 1973. Reciprocal averaging: An eigenvector method of ordination. *J Ecol* 61:237-251.
- Holm L, Sander C. 1994. Parser for protein folding units. *Proteins* 19:256-268.
- Islam SA, Luo J, Sternberg MJ. 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* 42:71-78.
- Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC. 1990. Atomic structure of the actin:dnase I complex. *Nature* 347:37-44.
- Lee BK, Richards FM. 1971. The interpretation of protein structures. Estimation of static accessibility. *J Mol Biol* 55:379-400.
- Lesk A M, Rose GD. 1981. Folding units in globular proteins. *Proc Natl Acad Sci USA* 78:4304-4308.
- Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. *J Mol Biol* 196:641-656.
- Moult J, Unger R. 1991. An analysis of protein folding pathways. *Biochemistry* 30:3816-3824.

- Murphy K, Bhakuni V, Vie D, Freire E. 1992. Molecular basis of cooperativity in protein folding. II. Structural identification of cooperative folding units and folding intermediates. *J Mol Biol* 227:293–306.
- Naor D, Fischer D, Jernigan RL, Wolfson HJ, Nussinov R. 1996. Amino acid pair interchanges at spatially conserved sites. *J Mol Biol* 256:924–938.
- Noguti T, Sasakibara H, Go M. 1993. Localization of hydrogen-bonds within modules in barnase. *Proteins* 16:357–363.
- Novokhatny VV, Kudinov SV, Privalov PL. 1994. Domains in human plasminogen. *J Mol Biol* 179:215–232.
- Nussinov R, Wolfson HJ. 1991. Efficient detection of motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA* 88:10495–10499.
- Otzen DE, Itzhaki LS, ElMasry NF, Jackson SE, Fersht AR. 1994. Structure of the transition state for the folding state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for the mechanism of protein folding. *Proc Natl Acad Sci USA* 91:10422–10425.
- Rose GD. 1979. Hierarchic organization of proteins. *J Mol Biol* 134:447–470.
- Privalov PL. 1996. Intermediate states in protein folding. *J Mol Biol* 258:707–725.
- Panachenko AR, Luthey-Schulten, Wolynes PG. 1996. Foldons, protein structural modules and exons. *Proc Natl Acad Sci USA* 93:2008–2013.
- Rashin AA. 1981. Location of domains in globular proteins. *Nature* 291:85–87.
- Segawa S, Richards FM. 1988. Identification of regions of potential flexibility in protein structures: Folding units and correlations with intron positions. *Biopolymers* 27:23–40.
- Shakhnovitch E, Abkevich V, Pitsyn O. 1996. Conserved residues and the mechanism of protein folding. *Nature* 379:96–98.
- Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. *J Mol Biol* 79:351–371.
- 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.
- Struthers MB, Cheng RP, Imperiali B. 1996. Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science* 271:342–345.
- Swindell MB. 1995. A procedure for detecting structural domains in proteins. *Protein Sci* 4:103–112.
- Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. 1996. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol* 260:604–620.
- Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. 1997. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Sci* 6:53–64.
- Wodak SJ, Janin J. 1981. Location of structural domains in proteins. *Biochemistry* 20:6544–6552.
- Wu LC, Grandori R, Carey J. 1994. Autonomous subdomains in protein folding. *Protein Sci* 3:359–371.
- Zehfus MH. 1993. Improved calculations of compactness and a reevaluation of continuous compact units. *Proteins* 16:293–300.
- Zehfus MH. 1994. Binary discontinuous compact protein domains. *Protein Eng* 7:335–340.
- Zehfus MH, Rose GD. 1986. Compact units in proteins. *Biochemistry* 25:5759–5765.

Appendix

Our cutting algorithm relies heavily on the calculation of the solvent accessible surface area (ASA), which is either buried in the interior or exposed on the surface. Since the evaluation of the ASA is time consuming, we have adopted the look-up table approach proposed by Zehfus (1993) to speed up the calculation. Cutting the 20 proteins listed in Table 1 took only about eight CPU minutes on a Silicon Graphics workstation (R4000 CPU and R4010 FPU).

The algorithm of Shrake and Rupley (1973) has been adopted in this work to calculate the ASA. The solvent accessible surface area of a protein is calculated numerically by summing the ASA represented by discrete spherical points. These are evenly distributed on a sphere with the radius of the van der Waals of the atom plus a probe ball of 1.4 Å in radius. In our implementation, starting with an icosahedron geometry, a series of 42, 162, and 642 spherical points were calculated with a quasi-uniform distribution. We have adopted the 162-point icosahedron to represent the surface of a sphere, since the difference between the surface area calculated by using 162 points and that calculated using 642 surface points is always within 0.1% for most of the proteins we have examined. If the distance between a spherical point of an expanded van der Waals atom A and a neighboring atom B is less than the sum of the van der Waals radius of atom B and the probe ball, the ASA of atom A is defined as buried by its neighboring atom B. In our calculation of the ASA, we exclude the ASA, which is buried by an atom belonging to the same residue or the two backbone atoms (C and N) that covalently link to that residue. The ASA of a residue can be buried by as many as seven neighboring residues.

As noted above, we have constructed a residue-based look-up table (Zehfus, 1993). For each residue in the table, the list of ASA buried by a list of residues is recorded. Since the ASA is also used to calculate the hydrophobicity of a folding unit, the list of ASA is separated into two lists, non-polar and polar. As the volume of a protein is evaluated numerically using the individual exposed ASAs, this procedure affords calculation of the volume of a cut unit with the same look-up table. For this reason, the volumes associated with each of the ASAs are also kept in the table. Once the look-up table is complete, it is straightforward and fast to calculate the four quantities determining the score of a cut folding unit.