Protein Science (1995), *4*:506-520. Cambridge University Press. Printed in the USA.

An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins

R. SOWDHAMINI AND TOM L. BLUNDELL

ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom

(RECEIVED September 16, 1994; ACCEPTED December 21, 1994)

Abstract

With a growing number of structures available in the Brookhaven Protein Data Bank, automatic methods for domain identification are required for the construction of databases. Domains are considered to be clusters of secondary structure elements. Thus, helices and strands are first clustered using intersecondary structural distances between C^{α} positions, and dendrograms based on this distance measure are used to identify domains. Individual domains are recognized by a disjoint factor, which enables the automatic identification and classification into disjoint, interacting, and conjoint domains. Application to a database of 83 protein families and 18 unique structures shows that the approach provides an effective delineation of boundaries and identifies those proteins that can be considered as a single domain. A quantitative estimate of the interaction between domains has been proposed. The database of protein domains is a useful tool for understanding protein folding, for recognizing protein folds, and for understanding structure-activity relationships.

Keywords: clustering of secondary structures; identification of domains; protein domains; protein three-dimensional structure; supersecondary structures

Even as the first structures were solved, proteins were found to have structurally distinct lobes (Phillips, 1966). However, the term "domain" was not assigned to compactly folded structures until later, when Wetlaufer (1973) recognized it to be a common feature in unrelated proteins. Since then, it has become clear that domains form an important level in the hierarchical organization of the three-dimensional structure of globular proteins, although not all proteins can be described as multidomain structures.

Domains are found in different combinations and often bring with them discrete functions. The relative disposition of domains in a protein may be important for the function and/or ligand binding (Lesk & Chothia, 1988; for a recent paper see Gerstein et al., 1994); for example, domain movements important to function have been described in aspartic proteinases (Sali et al., 1992) and in T4-lysozyme (Dixon et al., 1992). Domains of recently evolved proteins are frequently encoded by exons, reflecting gene fusion of simpler modules. Thus, domains are important both in protein folding and in biological function.

Reprint requests to: Tom L. Blundell, ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK; e-mail: ubcg91t@ccs.bbk.ac.uk.

The definition of a domain has undergone several stages of metamorphosis. A domain may be (1) functional (functional domain), for example, glutathione reductase or (2) defined on structural considerations (structural domain), for example, T4-lysozyme, or (3) an independent folding unit (folding domain), for example, γ-crystallin. Although each definition is valid when used in a particular context, often definitions are interchanged. We restrict the present analysis to structural domains. Although the identification of domains and their boundaries may often seem subjectively obvious, a general and automatic definition of domain boundaries is not straightforward. In particular, domains that are discontinuous (made up of more than one segment of the polypeptide chain in the amino acid sequence) or highly associated (characterized by a large number of contacts at the domain interface) are not easily defined.

Several algorithms for domain identification have been suggested and domains in proteins have been discussed by several groups (Rao & Rossmann, 1973; Liljas & Rossmann, 1974; Levitt & Chothia, 1976; Sternberg & Thornton, 1977; Wodak & Janin, 1981; Janin & Chothia, 1985; Zehfus & Rose, 1986; Kikuchi et al., 1988). In order to compile a database of protein structural domains, an automatic, fast, and reliable procedure is required.

Schulz (1977) proposed that domains have short distances between residues because they are structurally compact. The

reciprocal of the average distance between the C^{α} atom of a residue to other C^{α} atoms that are >7 residues and <25 residues away in the sequence was used to define domains; domain linkers obtain relatively low values. Go (1981) also exploited the fact that interdomain distances are normally larger than intradomain distances; all possible C^{α}-C^{α} distances were represented as diagonal plots (Go & Nosaka, 1987) in which there were distinct patterns for helices, extended strands, and combinations of secondary structures. However, both methods assumed that domains comprise an ideal compact structure of amino acid residues and this is not universal.

Rose (1979) considered a protein molecule as a rigid body and defined three mutually perpendicular axes passing through the centroid. The domain disclosing plane (defined by the larger two axes) and the cutting line (corresponding to the third axis) were employed to identify continuous chain segments that corresponded to compact domains. By measuring the entire protein volume and the volume of dissected segments, two segments A and B of a continuous polypeptide AB were said to be domains if they fell on either side of the domain disclosing plane. An error function was applied to every pair of polypeptide segments of the protein. An important prerequisite for the error function was the input of rough boundaries of the domains. This method works strictly only for continuous domains and may identify several substructures as individual domains.

A binary clustering algorithm (Crippen, 1978) considered proteins as several small segments that need not be the secondary structural components of the protein. The initial segments were clustered one after another based on intersegment distances. Segments with the lowest values were clustered and considered as a single segment thereafter. The stepwise clustering finally included the full protein.

Zehfus and Rose (1986) calculated compactness of substructures using solvent accessibility. Although this method can be extended to identify discontinuous domains (Zehfus, 1994), it is computationally expensive.

Argos (1990) analyzed protein domains in the context of the composition and conformation of domain linkers. A graphical inspection was used in order to identify domains and their boundaries. Holm and Sander (1994) used a contact matrix to group residues in a protein in order to identify domains. Islam et al. (1995) also use contact matrices for defining domains.

In this paper, we describe a method that subsumes the advantages of several other methods. C^{α}-C^{α} distances between secondary structures are represented in the form of average values termed "proximity indices" and the secondary structural organization is indicated in the form of dendrograms. Specific nodes in these dendrograms are identified as tertiary structural clusters of the protein; these include supersecondary structures and domains.

A ratio of the average proximity indices (ignoring interclusteral distances) to the average of all proximity indices, weighted for the aggregation of small subclusters and termed the disjoint factor, is employed as a discriminatory parameter to identify automatically clusters representing individual domains. The procedure is applied to independent protein structures that form representatives of a database of aligned homologous proteins (Overington et al., 1993) and single structures. It is shown that the domain boundaries usually match well with those reported by the authors in their crystal structure reports; but for some proteins, the procedure has identified convincing but hitherto unrecognized domains embedded in protein structures. It can also identify those proteins with single domain folds. It will be useful to quantify the extent of interaction between domains in order to classify them into disjoint, interacting, and conjoint types. Disjoint domains are those that have sparse domain interface, whereas conjoint domains are characterized by an elaborate domain interface and several contacts are present between the domain entities. Interacting domains are defined as those that contain intermediate degree of interactions between domains. A "disjoint factor" has been introduced that gives a quantitative measure of the extent of interaction. Accordingly, the calculated disjoint factor is used to classify domains into the three types.

Results and discussion

General distribution of proximity indices

The proximity index is a measure of the extent of interaction between pairs of secondary structures in a protein and is given by the average of all possible distances between C^{α} atoms of one secondary structure to the C^{α} atoms of the other. The distribution of proximity indices of secondary structural pairs in 20 independent proteins of varying size and fold (shown in Fig. 1) illustrates that the values tend to be smaller, as expected, where the two secondary structures are β-strands. However, for helix-helix interactions, the proximity indices are seldom less than 10 Å. Proximity indices of secondary structures belonging to different domains of a protein can be as high as 70 Å. By employing a clustering algorithm on these indices, secondary structures can be grouped into tertiary structural clusters representing supersecondary or tertiary structures.

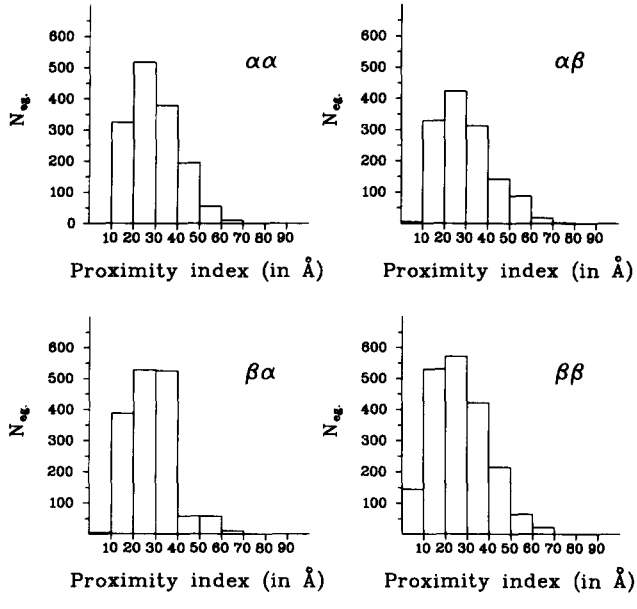

Fig. 1. Distribution of proximity indices between pairs of secondary structures (α: helix; β: extended strand) of 20 different proteins of varying sizes and folds. These 20 proteins form a subset of the 101 proteins used for analysis. $N_{eg.}$ is the number of examples.

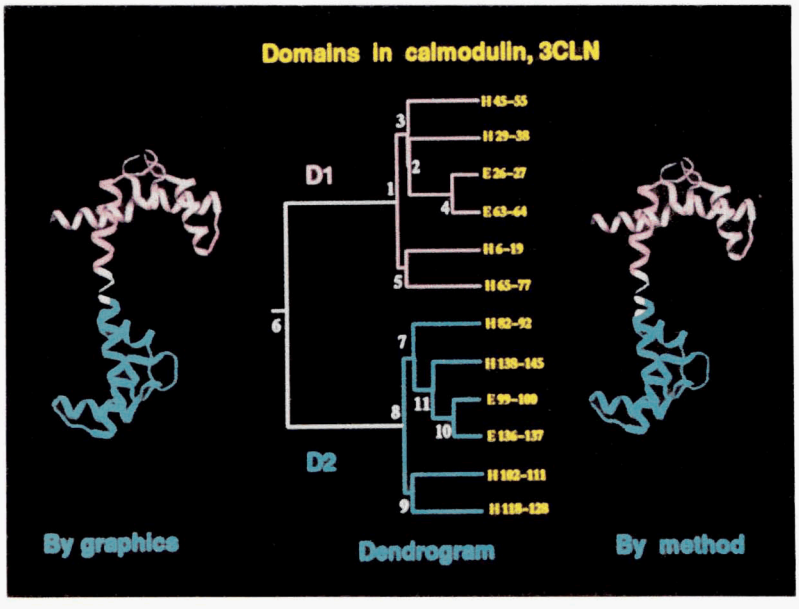

Fig. 2. Ribbon drawing of the crystal structure of calmodulin (PDB code, 3CLN) together with the secondary structural dendrogram. Different nodes are numbered and two clusters are prominent in the dendrogram. Domains identified by the authors as well from the dendrogram are compared. The N-domain is shown in pink and the C-domain in blue.

Five typical protein examples

Calmodulin

Calmodulin comprises a 148-residue polypeptide chain that folds into two distinct domains as revealed by the 2.2-Å crystal structure (Babu et al., 1988). Both domains have largely helical calcium-binding sites, the E-F hands, and the two domains are connected by an unusually exposed kinked helix. The domain boundaries have been identified as 5–78 and 82–147 by the authors. Figure 2 shows the C^{α} trace of calmodulin along with the secondary structural dendrogram; two prominent clusters (residues 6–77 and residues 82–145) in the dendrogram correspond to the two domains.

Porphobilinogen deaminase

Porphobilinogen deaminase (PDB code, 1PDA), a ubiquitous enzyme involved in the synthesis of hemes and chlorophylls, is composed of one discontinuous domain and two continuous domains (Louie et al., 1992) as shown in Figure 3. Domains 1 (residues 3–100; 201–218) and 2 (residues 106–194) are doubly wound, largely parallel β-sheets surrounded by helices. Domain 3 (222–307) is an open-faced three-stranded antiparallel β-sheet

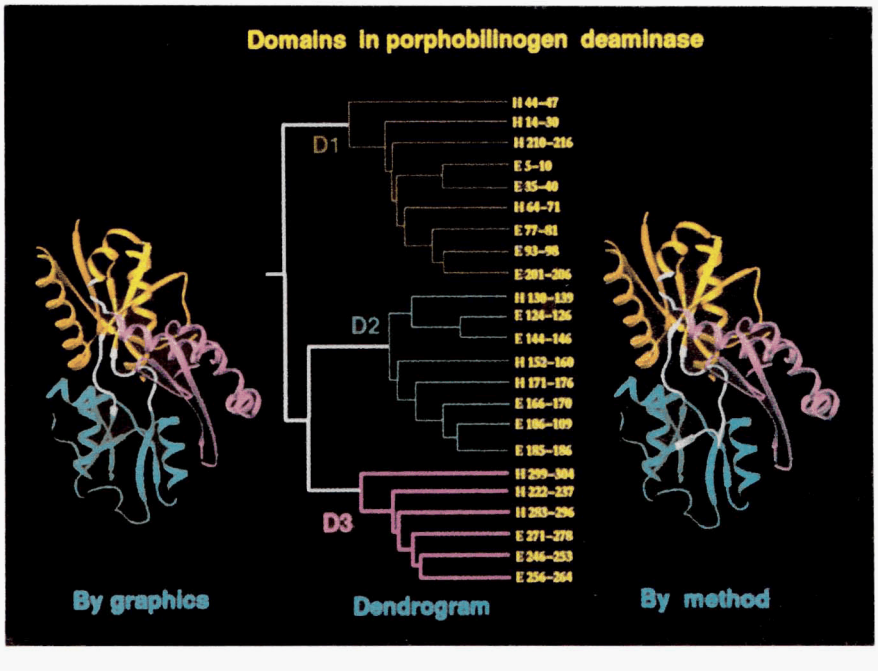

Fig. 3. Domains in porphobilinogen deaminase (PDB code, 1PDA). Three clusters are indicated by the tree diagram. Boundaries defined by the three clusters compare very well with the domain boundaries reported earlier (Louie et al., 1992) and examined by graphics (to the left).

with three helices on one side. These domain boundaries, reported by Louie et al. (1992), are confirmed by the boundary definitions delineated by the tree diagram (see Fig. 4) obtained by the clustering algorithm.

Endothiapepsin

The aspartic proteinases, for example endothiapepsin, are bilobal with the active site cleft between the two lobes. The structure (PDB code, 4APE) consists of five antiparallel β-sheets; there are two β-sheets in each of the N- and C-terminal domains and a further sheet (the central motif) at their interface (Blundell et al., 1990). From a subjective analysis of the interresidue contact matrices, Sali et al. (1992) conclude that the central motif should be considered separately. Figure 4 shows the C^{α} trace of the endothiapepsin structure and the dendrogram of the secondary structures. Three clusters can be identified from the dendrogram; residues 14–142 form one cluster, residues 0–6, 150–184A, and 311–325 form the second cluster, and residues 191–307 comprise the third cluster. Similar domain boundaries have been proposed by the authors while discussing domain flexibility.

Papain

The crystal structure of papain (Kamphuis et al., 1984) shows the protein has two discontinuous domains, defined by the authors by graphical inspection as: domain 1: 1–9; 112–207; domain 2: 10–111; 208–212.

Calculation of the intersecondary structural proximity indices followed by the clustering and construction of dendrograms also reveals the presence of two prominent clusters made up of discontinuous segments. An extended strand at residues 5–6 and secondary structures within residues 118–190 are grouped in cluster 1, whereas those within residues 25–112 and 207–210 are present in the second cluster of the dendrogram. Figure 5 shows the secondary structural dendrogram and C^{α} traces of papain with the two domains in different colors for boundaries defined by authors as well as by the present method.

Porin

Porin is a membrane-intrinsic β-protein comprised of 301 amino acids arranged as a 16-stranded tubular barrel with three short helices (Weiss & Schulz, 1992). Two prominent clusters are present in the secondary structural dendrogram of this protein: the N-terminal 32 residues and the C-terminal 58 residues together form the first cluster, which is separate from the rest of the β-barrel, which forms the second (Fig. 6). Porin is an unusual structure that resembles a hollow cylinder, and the two clusters roughly correspond to opposite faces of the cylinder. The clusters arise where β-strands in the barrel are long but interact with shorter β-strands. There are, nevertheless, still hydrogen bonds and close interactions between strands, so that it is not surprising that use of the disjoint factor (see below) indicates that these two clusters should be considered part of the same domains. Hydrogen bonding interactions are not specifically weighted while constructing the dendrograms. However, as discussed in the next section, we show that the present domain identification method does not consider these two clusters to represent individual domains, by means of a quantitative measure of the interactions between clusters.

Disjoint factor

Although clusters are evident from the dendrograms, we must address the following: (1) Of the various nodes and clusters in the tree, which are the ones that define domains? (2) How can we distinguish domains with secondary structures that interact with others? An example would be two domains that are separately folded but are hydrogen bonded through antiparallel β-strands. (3) How do we identify single domain folds that have

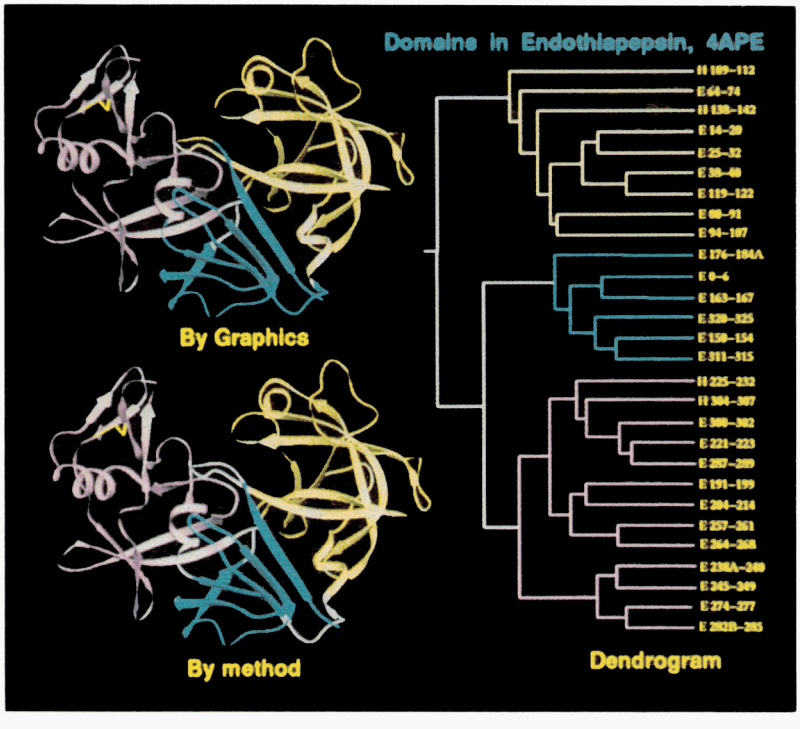

Fig. 4. Crystal structure of the aspartic protease, endothiapepsin (PDB code, 4APE; Blundell et al., 1990) together with the secondary structural dendrogram. Three clusters are evident from the dendrogram that correspond to the three noted by Sali et al. (1992). The N-domain is shown in blue, the central motif in yellow, and the C-domain in pink.

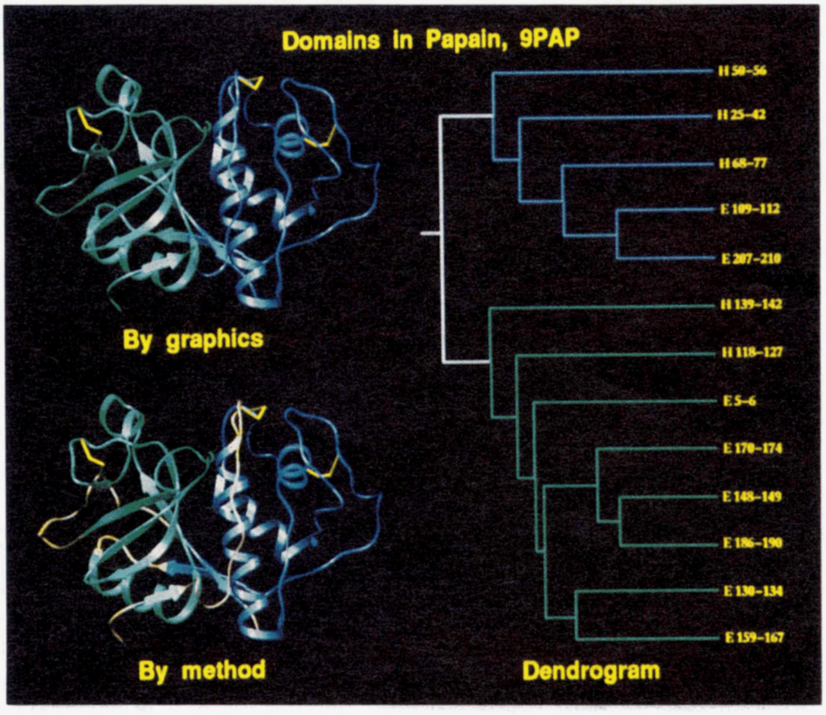

Fig. 5. Secondary structure-based ribbon representation of papain (PDB code, 9PAP; Kamphuis et al., 1984) and the tree diagram. Two domains proposed by the authors are colored differently and compared with the domains derived from the dendrogram.

multiple clusters in the dendrogram corresponding to supersecondary arrangements?

In order to assess these questions, a term called the "disjoint factor" is introduced. Once the dendrograms are constructed, the domain arrangement can be described by a situation that is a collection of nodes/clusters. For example, a node can define a pair of secondary structures. Clusters are higher order nodes and may represent more complex supersecondary structures or globular domains. For example, Figure 2 shows the dendrogram of calmodulin (3CLN) in which we have numbered every node (or branch) in the dendrogram. There are four pairs of secondary structures (node numbers 4, 5, 10, and 9). Together with four single secondary structures (H45–55, H29–38, H82–92, and H138–145), they give rise to two situations: situation 1: clusters 1

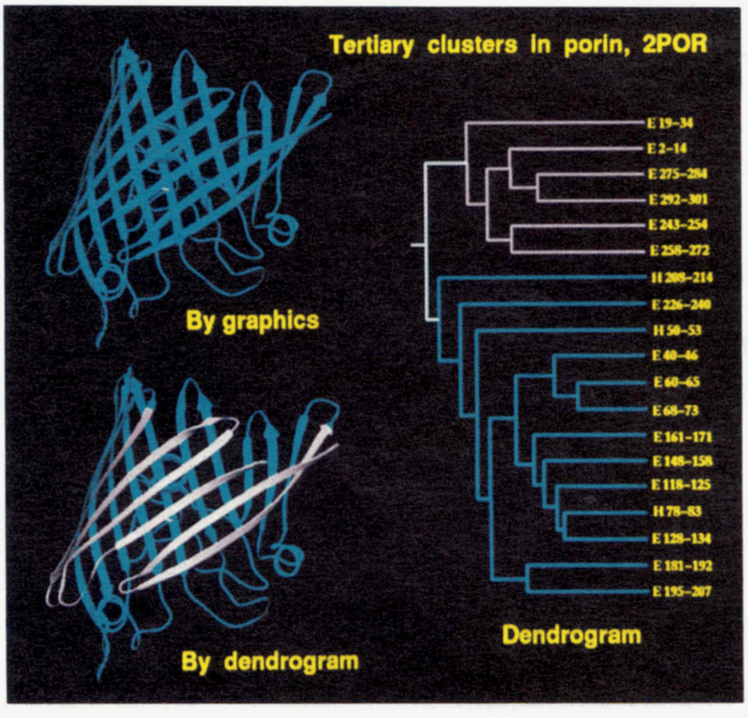

Fig. 6. Secondary structural dendrogram of porin (PDB code, 2POR) and the ribbon representation of the single domain fold (Weiss & Schulz, 1992). Two clusters are evident in the dendrogram; the cluster shown in pink comprises about 90 residues from the N- and C-terminal polypeptide segments. As shown by a low D_f value, this protein is considered as a single domain fold.

and 8 (two domains); situation 2: clusters 3, 5, 7, and 9 (four domains). For every situation, a "disjoint factor" (D_f) is calculated. This factor is related to a ratio that is the average proximity index value (ignoring indices of secondary structure pairs in different clusters) to the average proximity indices of all secondary structure pairs in a protein. This ratio, when weighted for close interactions between clusters (number of residues within 7 Å), is termed the "disjoint factor" (for a detailed definition of proximity index, cluster, situation, and disjoint factor, see Methods). The situation with the highest disjoint factor is considered to describe best the domain organization in a protein. Situations with D_f values lower than 1.0 do not represent domain arrangements.

Disjoint factors for the five examples and derivation of cut-offs for classification of domains

In the case of calmodulin, situation 1 (including the two domains) has a significantly better disjoint factor of 1.98 than situation 2 (D_f value is 1.35) involving four domains. However, the domains in calmodulin are near-ideal with no residues between the two clusters at a distance less than 7 Å. In the protein porphobilinogen deaminase (1PDA), the situation corresponding to the three clusters, described earlier, scores the highest D_f value of 1.64. The other two situations describe the domain organization by considering clusters of lower levels and have disjoint factors of 1.518 and 1.167. Similarly, D_f associated with the three-cluster situation in endothiapepsin (4APE) is 1.376 and the two clusters of papain (9PAP) obtain a D_f value of 1.06. Decreasing values of D_f indicate more closely interacting domains. In porin (2POR), the two tertiary structural clusters identified in the dendrogram have a calculated D_f value of less than 1.0 (0.726). Clearly, hydrogen bonding interactions between the strands have weighted down the value between the initially identified clusters and the method identifies porin to have a single domain.

A value of the parameter greater than 1 in general indicates clusters corresponding to domains. Increasing the value of D_f identifies domains that have decreasing interactions. This is also borne out by the fact that domains in papain (D_f is 1.06) are well defined but highly interacting, whereas domains in calmodulin do not interact. Forty other protein structures (independent of those in this analysis) were examined on the graphics and the domain boundaries suggested by the method were carefully compared. The correlation between the extent of interaction between domains and the calculated disjoint factor was also analyzed (data not shown). This led to the conclusion that we can define disjoint ($D_f > 1.5$) domains as in calmodulin and porphobilinogen deaminase, interacting ($1.25 \leq D_f \leq 1.5$) domains as in endothiapepsin or conjoint ($1.0 \leq D_f \leq 1.25$) domains as in papain.

Application of the method to representative structures from an alignment database and unique structures

The proteins used for the analysis (see Methods and Electronic Appendix for a list) include a total of 101 protein structures, 83 representative members belonging to different protein families (as defined by Overington et al., 1993) and 18 unique proteins. Domain boundaries suggested by the clustering of secondary structures were compared with those reported by the authors of the crystal structures in the literature and independently investigated using a graphical inspection. Table 1 shows a comparison of the boundary definitions by the procedure as well as those reported by the authors.

Domain boundaries have been identified well in most multidomain proteins, for example, 1LZ1, 3HLA, 3TLN, 1ATN, 3AAT, 6XIA, 2CD4, 2TBV-a, 2GCR, 2ALP, 3RP2, 4APE, 1FNR, 1FBP, 1MSB, 1RHD, 3LZM, 2TAA, 2TS1, 3GAP, 6ACN, and 6AT1 in Table 1. However, in other multidomain proteins, such as 1RNH, 2FB4lc, 3PHV, 1ACX, 2AZA, 2FB4lv, 1F3G, 1NSB, 1ACE, 1COL, and 3BCL, multiple clusters exist in the secondary structural dendrogram and the calculated D_f value is higher than one, indicating domains not identified by the authors.

Clusters with high D_f values often occur in small proteins that are either disulfide bonded or bound to metal ions (see Table 1, entries 4TGF, 1PK4, 1GSS, 2CTX, 6INS, 3B5C, 5PAL, 1YCC, 2CDV, and 2HHBa, which have disjoint factors greater than 1.0). They are small isolated segments, often stabilized by disulfides or a metal ion, that do not represent the true structural domains with hydrophobic interactions. For this reason, we do not consider clusters with less than 25 residues as domains.

Of the 101 proteins examined in this analysis, only in modified α1-antitrypsin (PDB code, 9API) were domains identified by a graphical inspection that were not identified by the procedure. Although two clusters exist in the dendrogram of 9API (not shown), the D_f value is significantly less than 1 (0.843). This protein is comprised of an extensive region of β-sheet composed of strands that are discontinuous in the sequence. A graphical inspection suggests that the protein can be separated into two domains mainly by the pattern of helical clusters that surround the central common β-sheet. The low value of D_f is understandable due to the presence of the β-sheet. Hence, this protein may be viewed as a single domain fold as implied by the procedure.

Isocitrate dehydrogenase (3ICD) has been described as three domains, a large $\alpha+\beta$ domain, one small α/β domain, and a smaller α/β clasp-like domain that is shared by two subunits of the functional dimer (Hurley et al., 1989). Our method splits the large $\alpha+\beta$ domain into two clusters, so isolating a helical subdomain that was noted by the authors.

In some proteins, for example 3BLM, 1ATN, 1ABP, 1PFK, 5LDH, 1GD1, 1COX, 1PII1, 5RUB, 1LAP, 2PHH, and 7CAT, the domain boundaries are somewhat different from those obtained by graphical inspection. In the N-domain of β-lactamase (3BLM), three discontinuous polypeptide segments are identified by the method, whereas the crystallographers propose only two discontinuous segments. This is due to the fact that the extra residues, corresponding to a helix at 201–213, closely interact with four other helices of the N-domain. A similar situation occurs in actin (1ATN) where the second domain, which is made up of a continuous polypeptide segment, is larger than that defined by the authors due to close interactions between helix 79–91 and the rest of the second domain. In the enzyme lactate hydrogenase, helix 244–263 is grouped with the N-domain due to its interaction with helix 31–43 of the N-domain. Similarly, close interactions of one or two secondary structure(s) with the rest of the residues in a domain, not easily noticeable on graphics, give rise to slightly different boundaries in the other proteins mentioned above.

In glutathione transferase (3GRS), four domains have been proposed by the authors on the basis of function. These com-

Table 1. *Domain boundaries of proteins used in the analysis*

Protein code	Domain boundary		Disjoint factor (D_f)	Classification
	By graphics	By method		
1NCP		No clear clusters	–	Single domain
1ZNF		No clear clusters	–	Single domain
1MRB		No clear clusters	–	Single domain
2MRB		No clear clusters	–	Single domain
1PPT		No clear clusters	–	Single domain
1BBL		No clear clusters	–	Single domain
6RXN		No clear clusters	–	Single domain
1PGX		No clear clusters	–	Single domain
2CI2		No clear clusters	–	Single domain
4FD1		Two clusters C1 2–4; 45–56 C2 25–34; 66–75	0.983	Single domain
1UTG		No clear clusters	–	Single domain
2HIP		No clear clusters	–	Single domain
1FXI		Two clusters C1 3–18 C2 48–88	0.290	Single domain
2ETI		No clear clusters	–	Single domain
4TGF		Two clusters C1 20–24; 29–33 C2 38–39; 45–46	2.253	Two isolated clusters
1BDS		No clear clusters	–	Single domain
1TAB		No clear clusters	–	Single domain
6INS		Three clusters C1 1A–16A; 2B–19B C2 24B–26B; 9D–26D C3 3C–16C	1.359	Two clusters
2OVO		No clear clusters	–	Single domain
5PTI		No clear clusters	–	Single domain
2CTX		Two clusters C1 2–13 C2 19–57	1.064	Two clusters
1PK4		Two clusters C1 72–74; 62–64 C2 1–2; 78–79	2.138	Two isolated clusters
1HDD		No clear clusters	–	Single domain
2CRO		Two clusters C1 2–12; 45–61 C2 17–35	0.838	Single domain
351C		No clear clusters	–	Single domain
3B5C		Two clusters C1 6–47; 75–79 C2 55–71	1.130	Two clusters
5PAL		Two clusters C1 8–32 C2 40–108	1.133	Two clusters
1YCC		Two clusters C1 3–13; 88–101 C2 50–74	1.344	Two clusters
2CDV		Two clusters C1 30–41 C2 9–20; 65–98	1.312	Two clusters
2HMQ		Two clusters C1 22–37; 91–104 C2 41–85	0.973	Single domain
1BP2		No clear clusters	–	
2CCY		No clear clusters	–	
2HHBa		Three clusters C1 4–17; 119–136 C2 21–71; 96–112 C3 76–89	1.217	Three clusters

(*continued*)

Table 1. (*Continued*)

Protein code	Domain boundary		Disjoint factor (D_f)	Classification
	By graphics	By method		
3CLN	Two domains D1 5-78 D2 82-147	Two clusters C1 6-77 C2 82-145	1.978	Disjoint domains
1ABM		Three clusters C1 20-28; 166-180 C2 30-80 C3 92-159; 186-196	1.285	Three clusters
1GSS		Two clusters C1 3-7; 28-73 C2 15-23; 185-197	1.797	Two isolated clusters
1PRC	Cytochrome subunit D1 1-36; 143-315 D2 37-142; 316-336	Four clusters C1 8-34; 244-247 C2 172-239 C3 52-81 C4 102-136; 262-309	1.602	Four domains
1LZ1	Two domains D1 1-39; 90-130 D2 40-88	Two clusters C1 5-36; 110-115 C2 43-100	1.250	Interacting domains
3HLA	Two domains D1 1-178 D2 185-262	Two clusters C1 3-179 C2 186-262	1.470	Interacting domains
9PAP	Two domains D1 1-9; 112-207 D2 10-111;208-212	Two clusters C1 5-6; 118-190 C2 25-112; 207-210	1.060	Conjoint domains
2CAB		Two clusters C1 32-50; 78-82; 108-124; 141-150; 191-212; 257-258 C2 56-70; 88-97; 131-134; 158-175; 216-226	0.043	Single domain
3BLM	Two domains D1 1-67; 168-256 D2 69-154	Two clusters C1 33-67; 180-193; 221-287 C2 72-154; 201-213	1.330	Interacting domains
3TLN	Two domains D1 1-157 D2 157-316	Two clusters C1 4-151 C2 159-312	1.451	Interacting domains
1ATN	Four subdomains D1 1-32; 70-144; 338-372 D2 33-69 D3 145-180; 270-337 D4 181-269	Four clusters C1 8-32; 103-136; 338-371 C2 35-91 C3 137-178; 274-330 C4 182-262	1.532	Disjoint domains
3ICD	Three domains D1 1-124; 318-416 D2 125-157; 203-317 D3 158-202	Four clusters C1 15-121; 326-366 C2 370-414 C3 126-154; 203-316 C4 164-197	1.469	Interacting domains
9API	Two domains D1 1-147 D2 148-393	Two clusters C1 23-79; 204-340; 363-388 C2 89-193; 344-357	0.843	Single domain
3AAT	Two domains D1 15-47; 326-410 D2 48-325	Two clusters C1 21-24; 313-405 C2 51-311	1.207	Conjoint domains
3GRS	Four domains D1 19-157 D2 158-293 D3 294-364 D4 365-478	Three clusters C1 19-50; 124-154; 326-354 C2 56-120; 158-216 C3 228-240; 369-461	1.703	Disjoint domains
3TRX	Two domains D1 1-72 D2 74-108	Two clusters C1 39-48; 95-104 C2 3-28; 53-90	1.16	Two clusters

(*continued*)

Table 1. (*Continued*)

Protein code	Domain boundary: By graphics	Domain boundary: By method	Disjoint factor (D_f)	Classification
1RNH		Two clusters C1 4-69; 115-141 C2 72-111	1.126	Conjoint domains
3FXN		No clear clusters	–	Single domain
1ETU		No clear clusters	–	Single domain
3DFR		Two clusters C1 2-32; 112-160 C2 38-105	1.244	Interacting domains
1GKY	Two domains D1 1-32; 82-186 D2 33-81	Three clusters C1 4-31; 82-121; 164-183 C2 125-157 C3 34-81	1.291	Interacting domains
1SBT		Two clusters C1 6-17; 175-201; 220-274 C2 27-152; 205-217	0.968	Single domain
3TMS		Two clusters C1 2-37; 213-220 C2 53-209; 229-250	0.665	Single domain
1ABP	Two domains D1 1-109; 255-284 D2 109-254; 285-291	Two clusters C1 7-81; 260-272 C2 111-252	1.566	Disjoint domains
1PFK	Two domains D1 1-160; 257-319 D2 162-252	Two clusters C1 3-123; 258-302 C2 139-246; 309-318	1.662	Disjoint domains
5LDH	Two domains D1 23-165 D2 183-331	Two clusters C1 24-161; 244-302 C2 166-237; 311-329	1.25	Interacting domains
1GD1	Two domains D1 1-149 D2 150-334	Two clusters C1 1-145; 315-330 C2 149-311	1.407	Interacting domains
2LIV	Two domains D1 1-110; 257-294 D2 111-256; 295-309	Two clusters C1 3-116; 259-316 C2 124-248; 330-342	1.563	Disjoint domains
1COX	Two domains D1 5-44; 226-316; 462-506 D2 45-225; 317-461	Two clusters C1 11-155; 232-303; 388-399; 444-505 C2 162-204; 324-380; 407-424	1.280	Interacting domains
1PII	Two domains D1 1-40; 84-162 D2 41-83; 163-221	Two clusters C1 39-54; 70-214; 232-235 C2 57-63; 220-226; 245-254	1.14	Conjoint domains
1TIM		No clear clusters	–	Single domain
1ALD		Two clusters C1 9-32; 73-301 C2 36-63; 303-337	1.066	Helical inserts in TIM barrel
1GOX		Two clusters C1 8-26; 165-205 C2 33-155; 227-356	1.080	Helical inserts in TIM barrel
6XIA	Two domains D1 10-321 D2 328-393	Two clusters C1 10-321 C2 323-383	1.298	Interacting domains
5RUB	Three domains D1 2-138 D2 142-420 D3 393-457	Three clusters C1 14-133; 292-315; 336-356 C2 143-287; 319-321; 364-369; 389-392 C3 376-383; 403-455	1.647	Disjoint domains
2FB4lc		Two clusters C1 116-120; 132-182; 193-208 C2 124-128; 184-189	1.078	Conjoint domains?
2CD4	Two domains D1 1-98 D2 99-173	Two clusters C1 2-102 C2 114-174	1.376	Interacting domains

(*continued*)

Table 1. (*Continued*)

Protein code	Domain boundary		Disjoint factor (D_f)	Classification
	By graphics	By method		
3PHV		Two clusters C1 10–24; 63–72 C2 32–59; 75–90	1.115	Conjoint domains?
1ACX		Two clusters C1 3–34; 50–63A; 89–93 C2 36–41; 67–82	1.421	Interacting domains
2AZA		Two clusters C1 4–9; 28–35; 92–98 C2 19–23; 49–52; 82–83; 108–128	1.405	Two clusters
2FB4lv		Two clusters C1 9–23; 61–75; 103–107 C2 33–47; 84–99	1.073	Conjoint domains?
1I1B		Two clusters C1 4–12; 42–62; 146–150 C2 17–29; 67–135	0.13	Single domain
2SOD		Two clusters C1 4–34; 93–99; 148–149 C2 39–87; 114–146	0.921	Single domain
1F3G		Two clusters C1 20–54; 136–167 C2 58–122	1.127	Conjoint domains?
1RBP		Two clusters C1 22–79; 166–167 C2 85–138	0.122	Single domain
2TBV-a	Two domains D1 106–270 D2 271–376	Two clusters C1 106–265 C2 274–376	1.500	Disjoint domains
2GCR	Two domains D1 1–81 D2 89–169	Two clusters C1 3–81 C2 89–169	1.475	Interacting domains
2ALP	Two domains D1 15A–107; 231–245 D2 132–228	Two clusters C1 15B–120K; 231–242 C2 135–230	1.079	Conjoint domains
3RP2	Two domains D1 16–21; 128–230 D2 28–122; 231–243	Two clusters C1 20–21; 135–230 C2 30–108; 231–242	1.183	Conjoint domains
4CNA		Two clusters C1 5–96; 140–175; 209–215 C2 103–130; 179–200	0.39	Single domain
4APE	Two domains with a central motif C 0–7; 150–184; 311–326 D1 8–143 D2 190–308	Three clusters C2 0–6; 150–184A; 311–325 C1 14–142 C3 191–307	1.38	Two domains with a central interacting motif
1NSB		Two clusters C1 91–97; 351–447 C2 112–315	1.13	Conjoint domains
4RHV	Four chains VP1 chain VP2 chain VP3 chain VP4 chain	Four clusters VP1 cluster VP2 cluster VP3 cluster VP1, VP3, and VP4 Interacting cluster		
1ACE		Four clusters C1 7–59; 96–146; 168–199 C2 151–155; 236–310 C3 79–82; 329–399; 518–532 C4 201–225; 319–324; 401–514	1.324	Interacting domains

(*continued*)

Table 1. (*Continued*)

Protein code	Domain boundary: By graphics	Domain boundary: By method	Disjoint factor (D_f)	Classification
1COL		Two clusters C1 8-67; 169-198 C2 76-164	1.074	Conjoint domains?
1FNR	Two domains D1 19-153 D2 160-314	Two clusters C1 38-151 C2 164-313	1.371	Interacting domains
1FBP	Two domains D1 6-199 D2 201-225	Two clusters C1 13-199 C2 209-334	1.277	Interacting domains
1LAP	Two domains D1 1-150 D2 151-484	Two clusters C1 3-68; 84-145 C2 76-83; 151-483	1.247	Conjoint domains
1MSB	Two subdomains D1 107-152 D2 158-212	Two clusters C1 107-135; 213-220 C2 155-207	1.040	Conjoint domains?
1PGD	Two domains D1 1-172 D2 178-433	Three clusters C1 5-175; 354-381 C2 178-291; 315-348; 392-432 C3 439-464	1.551	Disjoint domains
1RHD	Two domains D1 1-157 D2 158-293	Two clusters C1 9-136 C2 161-281	1.398	Interacting domains
3LZM	Two domains D1 1-72 D2 74-164	Two clusters C1 3-80 C2 85-155	1.327	Interacting domains
2PAB	Single domain	No clear clusters	–	Single domain
2PHH	Two domains D1 1-70; 100-184; 269-345 D2 70-99; 185-268; 348-386	Two clusters C1 5-40; 102-157; 277-283; 298-318 C2 47-99; 175-274; 289-290; 328-385	1.343	Interacting domains
2TAA	Three domains D1 1-121; 180-350 D2 121-179 D3 355-478	Three clusters C1 11-118; 182-375 C2 125-173 C3 388-464	1.426	Interacting domains
2TS1	Two domains D1 1-222 D2 224-319	Two clusters C1 2-220 C2 248-318	1.242	Conjoint domains
3BCL		Two clusters C1 4-68; 122-127; 244-279; 300-357 C2 71-117; 136-229; 284-290	1.185	Conjoint domains?
3GAP	Two domains D1 1-131 D2 134-208	Two clusters C1 9-134 C2 139-189	1.333	Interacting domains
6ACN	Three domains D1 1-310 D2 335-495 D3 550-750	Three clusters C1 18-314; 517-518 C2 337-495 C3 552-750	1.543	Disjoint domains
6AT1	*Catalytic chain* Two domains D1 17-129; 285-304 D2 135-279 *Regulatory chain* Two domains D1 8-98 D2 102-153	Two clusters C1 1-137; 293-310 C2 138-292 Two clusters C1 15-95 C2 102-150	1.419 1.342	Interacting domains Interacting domains
7CAT	Two domains D1 69-149; 212-365 D2 153-204; 434-499	Two clusters C1 10-16; 77-147; 213-237; 259-364 C2 160-198; 246-255; 440-499	1.321	Interacting domains

prise residues 19–157 of the FAD binding domain, residues 158–293 of the NADPH domain, residues 294–364 of the central domain, and residues 365–478 of an interfacial domain. The protein, which is functional only as a symmetrical dimer (Karplus & Schulz, 1987), has two symmetric active sites with protomers contributing to each active site. The present procedure identifies three structural domains, which, as expected, do not have the same boundaries as those of the classical domains. As shown in Figure 7, one of them is comprised of three discontinuous polypeptide segments (19–50; 124–154 and 326–354), whereas the other two are made up of two discontinuous polypeptide segments (56–120; 158–293 and 228–240; 369–461).

The nucleotide kinases have two domains, one of which is similar in both adenylate and guanylate kinases, whereas the other "insert" domain differs between the two proteins. The three clusters in the secondary structural dendrogram of guanylate kinase (1GKY) correspond to these two domains and a cluster of two roughly antiparallel helices. The third cluster corresponds to a highly flexible "flap" or "lid" domain in the adenylate kinases (1AKE and 1AK3) that undergoes large movements upon AMP binding (Berry et al., 1994) not present in the guanylate kinases. Indeed, the secondary structural dendrograms of adenylate kinases have three clusters representing the three domains.

In catabolite gene activator (protein code, 3GAP), the protein chain is folded into two domains that are connected by a long helix. One of the domains makes a significant number of contacts with the central helix. The exclusion of the central helix (residues 113 and 124) from either of the two domains gives a higher D_f than a situation where the central helix is considered part of one domain. A better representation would be to split the central helix into two segments with each part of a domain, analogous to the domain representation in calmodulin (see Fig. 2), where there is a natural kink in the central linking helix. Thus, once the tertiary clusters describing domains are identified, long secondary structures in domain linkers and loop regions *between* clusters may be examined specifically for local interactions and partial secondary structures may in this case be allocated to domains.

Fig. 7. Domains in glutathione transferase (PDB code, 3GRS). Three prominent clusters are noted in the secondary structural dendrogram (not shown). Structural domains identified here do not correspond to the functional domains defined by the authors (Karplus & Schulz, 1987). The first six residues of the helix 101–120, which link two of the domains (shown in green and pink), are colored in green for the sake of clarity. Monomer coordinates were used in domain identification.

In 6-phosphogluconate dehydrogenase (PDB code 1PGD), the C-terminal 25 residues (residues 439–464) form a separate cluster in the secondary structural dendrogram. Whether this region, which is comprised of a helix and two antiparallel β-strands, can be considered as a separate domain or not is questionable.

When D_f is less than 1.0, the domain organization attributed to the protein is definitely unfavorable. Several proteins in the database that are single domain folds have two clusters in the dendrogram with D_f values less than 1.0. Examples include 4FD1, 1FXI, 2CRO, 2HMQ, 2CAB, 3TMS, 1I1B, 2SOD, 1RBP, 4CNA, 2PAB, and 2POR. Interactions between the clusters allow the disjoint factor to demarcate true domains from those that involve intimately interacting secondary structural clusters. However, those proteins that have D_f values in the range 1.0–1.1 (for example, 2FB4lc, 3PHV, 2FB4lv, 1F3G, 1COL, 1MSB, and 3BCL) have doubtful domain organizational situations.

In the 101 proteins used for the present analysis, as shown in Table 1, domains have been classified into one of the three categories: Ten proteins have disjoint factors greater than 1.5 and hence have disjoint domains. Figure 8A shows the C^{α} trace of phosphofructokinase (1PFK, D_f value is 1.66), which has two disjoint domains. In 24 proteins (about 25%), the domains are of the interacting type. Figure 8B shows the C^{α} trace of an example of a pair of interacting domains in 2PHH. Bilobal proteins with elaborate domain interfaces have low disjoint factors. In the present analysis, 17 proteins have conjoint domains. An example of conjoint domains is the mammalian serine protease, rat mast protease (3RP2), shown in Figure 8C. Although such a quantitative estimate has enabled an interaction-based classification of domains, clearly, borderline cases are to be treated with caution.

Using an extended data set that includes 61 independent structure determinations apart from all members in the alignment database, the domain identification algorithm has been used to identify 581 domains in 447 protein structures. A detailed analysis of the domains of this extended dataset, in terms of secondary structure and fold similarities, the construction of domain templates, and their application to fold recognition are reported elsewhere (R. Sowdhamini, S. Rufino, M.S. Johnson, & T.L. Blundell, manuscript in prep.).

Conclusion

The method described above uses a clustering algorithm in order to represent the organization of secondary structures into three-dimensional domains. Domain organization is explicit from the tree diagrams, and it is possible to identify domain boundaries even for complicated situations.

The procedure presupposes domains to be compact folding units such that the proximity indices of secondary structures between domains would be significantly higher than those within

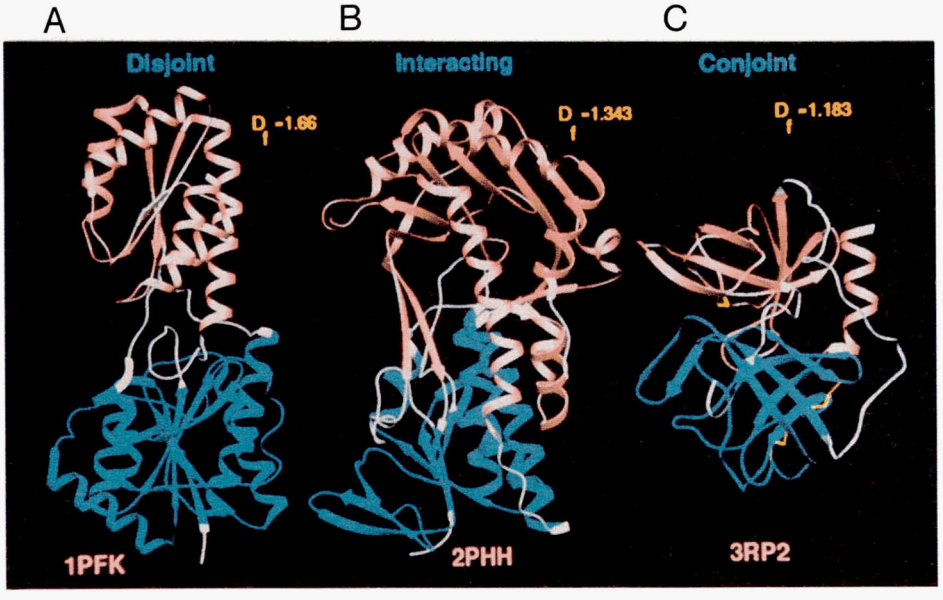

Fig. 8. Illustrative examples of pairs of disjoint, interacting, and conjoint domains in three different protein structures: domains in (**A**) phosphofructokinase, (**B**) *p*-hydroxybenzoate hydroxylase, and (**C**) rat mast cell protease. PDB codes and D_f values are marked.

a domain. For ideal compact domains, the average proximity index ignoring interdomain indices would be higher in magnitude than the average proximity index where all distances are considered. However, this compactness is not made as a rigid entity and certain allowance for interactions between domains is also permitted in this method. This procedure provides a versatile approach to domain identification, in which a reliable quantitative estimate of the extent of interaction between domains is proposed. The classification of domains thenceforth into three types, disjoint, interacting, and conjoint, provides a useful guide to the understanding of protein structure and function. The definition of protein domains involves a degree of subjectivity, often guided by an operational requirement such as to understand protein folding, evolution, function, and so on. There is no absolute definition of a protein domain and therefore it is difficult to assess methods quantitatively in a more rigorous fashion.

One feature of the present analysis is that only secondary structures are considered in distance matrices. Because the contacts involving loop regions are not represented, this obviously leads to an approximate representation. Additionally, considering only the secondary structures implies that loops that act as domain linkers will not be included.

As illustrated in the catabolite gene activator (PDB code, 3GAP), long secondary structures in linker regions may be assigned to one of the identified domains and may still maintain appreciable contacts with the rest of the domains. Using the present version of the method, it is not possible to assign automatically smaller segments of long secondary structures to subdivide the interactions of the long secondary structures in domain boundaries and assign smaller segments to individual domains. Thus, long secondary structures in domain linkers may be segmented to smaller regions and specific interactions be analyzed for contact studies.

This domain identification algorithm has been coded as a computer program DIAL. The list of domains in the present database (Table 1) is also available from the authors on request. We will be updating the domain database periodically. A more detailed analysis on 447 proteins uses structural comparison algorithm (Rufino & Blundell, 1994) to cluster protein domains and to derive domain templates useful for fold recognition methods (R. Sowdhamini, S. Rufino, M.S. Johnson, & T.L. Blundell, manuscript in prep.).

Methods

Tree construction

Secondary structures were identified on the basis of main chain hydrogen bonding patterns, using the program SSTRUC (D. Smith, unpubl. results), which implements the algorithm of Kabsch and Sander (1983) to define regions of α-helix and β-strand. A proximity index, p, was associated with every pair of secondary structures.

$$p_{i,j} = \frac{\sum_{k=1}^{k=n_i} \sum_{l=1}^{l=n_j} d_{k,l}}{n_i \times n_j}$$

where n_i refers to the number of residues in the secondary structure i, n_j refers to the number of residues in the secondary structure j, and d_{kl} is the distance between the C^α atom of kth residue of secondary structure "i" and the C^α atom of the lth residue of secondary structure "j."

The calculated proximity indices were used to perform clustering using KITSCH, which is a part of the phylogeny inference package PHYLIP (Felsenstein, 1985). The program involves bootstrap sampling and computes the "best" tree chosen by the parsimony and compatibility methods. This leads to a majority-rule consensus tree that has higher confidence limits and better statistical significance. This is a modified version of the original Fitch–Margoliash method (Fitch & Margoliash, 1967), where local swapping of the phylogeny branches has been included

to enable exploration of numerous topologies. The results of clustering analysis were transformed into a format suitable for drawing dendrograms using DRAWTREE (Z.-Y. Zhu, unpubl. results).

Identification of clusters from dendrograms

Starting from the secondary structural elements, nodes may represent even a pair of secondary structures while clusters are higher order nodes. Whereas nodes can be as small as four residues (a pair of two-residue extended strands), clusters with a size smaller than 25 residues are generally not considered. Although the choice of a lower limit for the size of the cluster seems arbitrary, it is well known that domains are generally between 40 and 200 residues in length (Wetlaufer, 1973). From a given dendrogram describing the secondary structural organization of a protein, it is possible to derive more than one combination of cluster. Each combination of possible clusters is referred to as a "situation." Out of several possible situations, only one is best suited to describe the domain organization of a protein.

Parameter for assessing compactness of domains and automatic identification of domain boundaries

Let us consider a "situation" where n_s clusters are identified from the secondary structural dendrogram of a protein. The disjoint factor (D_f) gives a measure of the physical interaction between identified clusters by comparing the mean proximity indices of secondary structures within clusters with the mean proximity indices of all secondary structures. $D_f = \alpha \times W_{1,2} \times W_{1,3} \cdots W_{n_s-1,n_s}$, where α is a ratio given by:

$$\alpha = \frac{\dfrac{\sum_{i=1}^{i=nt-1} \sum_{j=i+1}^{j=nt} p_{i,j}}{\dfrac{nt(nt-1)}{2}}}{\dfrac{\sum_{k=1}^{k=n_s} \sum_{ii=1}^{ii=ist(k)-1} \sum_{jj=ii+1}^{jj=ist(k)} p_{ii;k,jj;k}}{\dfrac{ist(k)(ist(k)-1)}{2}}}$$

where nt = total number of secondary structures in the protein; $p_{i,j}$ = proximity index between secondary structures i and j; $ist(k)$ = number of secondary structures in cluster (k); and W is a weight attached to pairs of clusters.

The weighting promoted the identification of clusters that are reasonably large and aggregates of highly interacting clusters (for example, residues that are part of a β-sheet may initially be part of two different clusters).

$$W_{1,2} = \frac{\sum_{i=1;1}^{i=1;l} \sum_{j=2;1}^{j=2;m} n(i) \times n(j) - \sum_{i=1;1}^{i=1;l} \sum_{j=2;1}^{j=2;m} \sum_{ii=n(i)}^{ii=n(i)} \sum_{jj=n(j)}^{jj=n(j)} d^2_{1;i;ii,2;j;jj}}{\sum_{i=1;1}^{i=1;l} \sum_{j=2,1}^{j=2,m} n(i) \times n(j)}$$

l = number of secondary structures in cluster 1; m = number of secondary structures in cluster 2; $n(i)$ = number of residues in secondary structure (i); $n(j)$ = number of residues in secondary structure (j); $d_{1;i;ii,2;j,jj}$ = number of residues within 7.0 Å between secondary structure i of cluster 1 and secondary structure j of cluster 2.

This weighting lowers the value only in the case of small clusters and does not significantly affect the values when the clusters do not have any interactions and when the clusters have minimal interactions at the interface.

Proteins used for the analysis

Coordinates of protein structures were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977). Proteins used for the present analysis include representative members from 86 protein families and 20 unique structures and have domains with different levels of interaction and some proteins with single domain folds. The Electronic Appendix contains a complete list of the protein names along with the family they belong to and the source. Domain boundaries defined by the present procedure have been compared with those proposed by the crystallographers and also independently confirmed by graphics. The domain boundaries are listed in Table 1. The ribbon representation of protein structures was performed using the software SETOR (Evans, 1993).

Acknowledgments

This work is funded by the Imperial Cancer Research Fund. We thank Dr. Zhan-Yang Zhu for providing the program DRAWTREE. We thank Dr. Mike Sternberg and his colleagues for sending us a preprint of their paper.

References

Argos P. 1990. An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. *J Mol Biol 211*:943–958.

Babu YS, Bugg CE, Cook WJ. 1988. Structure of calmodulin refined at 2.2 Å resolution. *J Mol Biol 204*:2013–2018.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Berry MB, Meador B, Bilderback T, Liang P, Glaser M, Phillips GN Jr. 1994. The closed conformation of a highly flexible protein: The structure of *E. coli* adenylate kinase with bound AMP and AMPPNP. *Proteins Struct Funct Genet 19*:183–198.

Blundell TL, Jenkins JA, Sewell BT, Pearl LH, Cooper JB, Wood SP, Veerapandian B. 1990. X-ray analyses of aspartic proteinases. The three-dimensional structure at 2.1 Å resolution of endothiapepsin. *J Mol Biol 211*:919–941.

Crippen GM. 1978. The tree structural organisation of proteins. *J Mol Biol 126*:315–332.

Dixon MM, Nicholson H, Shewchuk L, Baase WA, Matthews BW. 1992. Structure of a hinge-bending bacteriophage-T4 lysozyme mutant, Ile 2→ Pro. *J Mol Biol 227*:917–933.

Evans SV. 1993. SETOR – Harware-lighted 3-dimensional solid model representations of macromolecules. *J Mol Graphics 11*:134–138.

Felsenstein J. 1985. Confidence-limits on phylogenies – An approach using the bootstrap. *Evolution 39*:783–791.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science 15*:279–284.

Gerstein M, Lesk AM, Chothia C. 1994. Structural mechanisms for domain movements in proteins. *Biochemistry 33*:6739–6749.

Go M. 1981. Correlation of DNA exonic regions with protein structural units in hemoglobin. *Nature 291*:90–92.

Go M, Nosaka M. 1987. Protein architecture and the origin of introns. *Cold Spring Harbor Symp Quant Biol LII*:915–924.

Holm L, Sander C. 1994. Parser for protein folding units. *Proteins Struct Funct Genet 19*:256–268.

Hurley JH, Thorsness PE, Ramalingam V, Helmers NH, Koshland DE, Stroud RM. 1989. Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase. *Proc Natl Acad Sci USA 86*:8634–8639.

Islam SA, Lud J, Sternberg JE. 1995. Identification and analysis of domains in proteins. *Protein Eng.* Forthcoming.

Janin J, Chothia C. 1985. Domains in proteins – Definitions, location and structural principles. *Methods Enzymol 115*:420–430.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure – Pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kamphuis IG, Kalk KH, Swarte MBA, Drenth J. 1984. Structure of papain refined at 1.65 Å resolution. *J Mol Biol 179*:233–256.

Karplus PA, Schulz GE. 1987. Refined structure of glutathione reductase at 1.54 Å resolution. *J Mol Biol 195*:701–729.

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

Lesk AM, Chothia C. 1988. Elbow motion in the immunoglobins involve a molecular ball-and-socket joint. *Nature 335*:188–190.

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature 261*:552–558.

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

Louie GV, Brownlie PD, Lambert R, Cooper JB, Wood SP, Blundell TL, Warren MJ, Woodcock SC, Jordan PM. 1992. Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site. *Nature 359*:33–39.

Overington JP, Zhu ZY, Sali A, Johnson MS, Sowdhamini R, Louie GV, Blundell TL. 1993. Molecular recognition in protein families – A database of aligned 3-dimensional structures of related proteins. *Biochem Soc Trans 21*:597–604.

Phillips D. 1966. The three-dimensional structure of an enzyme molecule. *Sci Am Nov*:78–90.

Rao S, Rossmann M. 1973. Comparison of super-secondary structures in proteins. *J Mol Biol 76*:241–256.

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

Rufino S, Blundell TL. 1994. Structure-based identification and clustering of protein families and superfamilies. *J Comput Aided Mol Design 8*: 5–27.

Sali A, Veerapandian B, Cooper JB, Moss DS, Hofmann T, Blundell TL. 1992. Domain flexibility in aspartic proteinases. *Proteins Struct Funct Genet 12*:158–170.

Schulz GE. 1977. Structural rules for globular proteins. *Angew Chem Int Ed 16*:23–33.

Sternberg MJE, Thornton JM. 1977. On the conformation of proteins: Towards the prediction of strand arrangements in β-pleated sheets. *J Mol Biol 113*:401–418.

Weiss MS, Schulz GE. 1992. Structure of porin refined at 1.8 angstroms resolution. *J Mol Biol 227*:493–509.

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

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

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.