Location of Structural Domains in Proteins†

Shoshana J. Wodak and Joël Janin*

ABSTRACT: We use surface area measurements based on atomic positions to give a quantitative definition of structural domains in proteins. Segments of the polypeptide chain making a minimum of interactions with the rest of the protein structure are identified on *interface area scans*, where the area B of the interface between a N-terminal segment of i residues and the complementary C-terminal segment is plotted as a function of i. Domain boundaries appear as minima of B in the scans. The procedure may be iterated to build a hierarchy of subdomains. It detects only continuous domains made of a single stretch of polypeptide chain but may be extended to detect such domains in the presence of discontinuous ones. Domains defined from interface area scans fit very well with globular structural regions identified by inspection of protein models [Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697–701]. They do not in general correspond to the repeated structural units observed in some proteins by superposition studies. In hemoglobin and hen lysozyme, the domains do not correspond to the coding sequences separated by introns in the genes.

Inspection of molecular models derived from X-ray studies shows that in all but the smallest proteins, the polypeptide chain folds into several globular units, sometimes loosely connected (Wetlaufer, 1973). These units are commonly called domains, or *structural domains* to indicate that they are identified from the three-dimensional structure. Biochemical experiments such as limited proteolysis are another way of defining domains as fragments that may be stable and refold after denaturation in the absence of the rest of the chain. Domains often carry out specific functions, and it has been suggested that they are the units of protein evolution as well as protein structure (Edelman et al., 1969; Rossmann et al., 1974); multidomain proteins performing complex functions would result from the fusion of genes coding for each domain.

Location of structural domains by visual inspection of models suffers from subjectivity. Objective definitions and algorithms to find domains from atomic positions are therefore useful (Crippen, 1978; Rose, 1979). We use here surface area criteria to detect structural domains as regions of the protein structure where most of the interactions between atoms or residues occur within the regions and least without. We apply to several proteins an algorithm which finds limits of structural domains made of a single stretch of polypeptide chain (continuous domains). The results are in good agreement with previous descriptions of protein structures on the basis of visual inspection. In particular, we confirm the presence of two continuous domains in globins and of two domains in hen lysozyme, one of which is discontinuous. They do not correlate with the DNA coding sequences or exons in the hemoglobin and lysozyme genes.

Methods

Atomic Coordinates and Surface Area Calculations. Atomic coordinates are obtained from the Cambridge Protein Data Bank (Bernstein et al., 1977), except for glycogen phosphorylase, gift of R. Fletterick. The area of the protein surface in contact with the solvent, its accessible surface area, may be computed from atomic positions by using the geometrical algorithm of Lee & Richards (1971) implemented into a computer program of M. Levitt or by an analytical approximation, which expresses accessible surface areas as a function of distances between pairs of atoms (Wodak & Janin, 1980). The analytical approximation is several orders of magnitude faster to compute and better adapted to repetitive calculations. We apply it to simplified models of protein structures (Levitt, 1976), where each amino acid residue is represented by a single sphere centered on the center of mass of its side chain (C_α included). Sphere radii for each type of residue are taken from Wodak & Janin (1978). The simplified model provides a further gain in computational speed, at little cost in precision, as the surface area calculations are rather insensitive to changes in atomic positions no larger than the radius of the solvent probe (1.4 Å).

Calculations of surface areas with the analytical approximation and simplified protein models are comparable in speed to C_α distance measurements used in diagonal plots (Rossmann & Liljas, 1974). The values of the accessible surface areas are in good agreement with those given by the Lee & Richards procedure applied to detailed atomic models: on a sample of 21 proteins, including those discussed below, total accessible surface areas calculated in either way differ by an average of 1.4%. None differs by more than 6%.

Interface Area Scans. The interface area B, that is, the *surface area buried in contacts between two groups of atoms or residues*, is defined as

$$B = A_1 + A_2 - A_{12} \quad (1)$$

where A_1 and A_2 are the accessible surface areas of each of the two groups alone and A_{12}, the accessible surface area of the two together.

Interface areas define domains in the following manner. The polypeptide chain is assumed to be cleaved after residue i while its conformation is retained. The interface area of the N-terminal segment (1 to i) with the C-terminal segment (i + 1 to the last residue) is calculated. The cleavage point is moved along the chain, yielding what we call an interface area scan. On this scan, *putative limits of domains show as minima of the interface area*: at these points, cleavage creates two artificial subunits which have a minimum of contacts and most closely resemble real protein subunits.

Since the interface area of the two segments depends on their size, B drops sharply to zero at both ends of the scans, which tend to be bell shaped as in Figure 1. Such bell-shaped scans are characteristic of single domain proteins; other scans

† From the Laboratoire de Chimie Biologique, Université Libre de Bruxelles, Rhode-St-Genèse, 1640, Belgium (S.J.W.), and Unité de Biochimie Cellulaire, Département de Biochimie et Génétique Moléculaire, Institut Pasteur, 75724 Paris, Cedex 15, France (J.J.).
Received April 29, 1981.

0006-2960/81/0420-6544$01.25/0

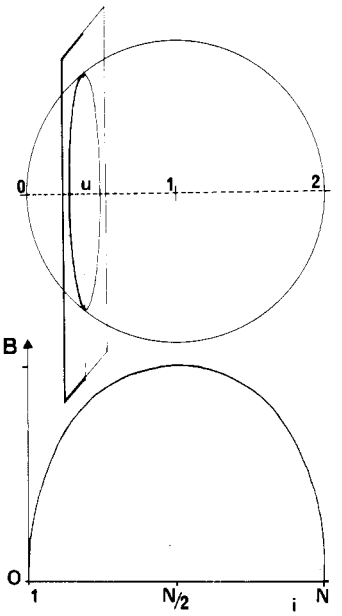

FIGURE 1: Interface area scan of a sphere. A sphere, representing the protein molecule is cleaved into two fragments by a vertical plane. The N-terminal fragment has i residues and the C-terminal, $N - i$; i and $N - i$ are proportional to the volumes of the spherical caps on the left and right of the dividing plane. The interface area B is twice the area of the circle of intersection. When the abscissa u of the plane increases from 0 to 2, the plane moves through the sphere from left to right; $i = (N/4)u^2(3 - u)$ increases from 0 to N, and $B = (A/2)u(2 - u)$ goes through a maximum equal to half the total area A of the sphere. The bell-shaped curve representing B as a function of i may also be taken to represent the interface area scan of a globular protein.

contain deep minima which define domain boundaries.

Iterative Cleavage Algorithm. Interface area scans may be used to construct a tree of binary divisions of the polypeptide chain, representing a hierarchy of domains and subdomains in the protein (Rose, 1979). We first cleave the chain at the lowest significant minimum of B in an interface area scan run on the complete protein. Then we run similar scans on each of the two fragments, find the lowest significant minima, and cleave the fragments. The process may be repeated for as long as the scans contain significant minima. A minimum at position i will be considered as significant if B is at least 2σ lower at position i than its highest values in segment 1 to i and segment $(i + 1)$ to N (end of the chain). In the same way, minima differing by 2σ or less are considered as equivalent.

The noise level σ is estimated from differences in interface areas at adjacent positions i and $i + 1$ along the scans:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N-1}(B_{i+1} - B_i)^2 \qquad (2)$$

σ is close to 100 Å^2 for all proteins examined.

An application of the iterative cleavage algorithm to glycogen phosphorylase is illustrated below. In this as in most cases that we tried, minima of B in scans of the second or third generation are also minima of B in the primary scan run on the complete protein. Thus, most of the information on subdomain structure that the algorithm yields is present in the primary scan. When the scan contains several minima, iterative cleavage may still be used to rank them.

Two-Dimensional Interface Area Scans. Interface area scans cleave protein structures into domains formed by continuous stretches of polypeptide chain (continuous domains). They might not apply to proteins where discontinuous domains are interrupted by continuous ones, as the chain must be cleaved at more than one point at a time to define the domains. Such cases may still be treated by measuring interface areas as follows. A segment of variable length L is moved along the polypeptide chain, and the area B of its interface with the rest of the protein chain, now composed of two segments, is calculated. B is then plotted in two dimensions, as a function of the length L and of the position i of the N terminus of the chain segment (see Figure 6 below). Chain segments making relatively few external contacts for their size appear as minima of B for given values of L and i. Segments (i to $i + L - 1$) corresponding to these minima qualify as domains according to our definition.

We applied this procedure to all proteins discussed here, save glycogen phosphorylase. We observe that with few exceptions, minima of B in two-dimensional plots correspond to minima of B in interface area scans. Thus, two-dimensional searches, which require much heavier computations, are useful only when the scans are ambigous, as in the case of hen lysozyme discussed below.

Globularity Index. Domains corresponding to minima of B have a minimum of contacts with the rest of the protein. The procedure gives no indication on their eventual stability as isolated structures. Compact globular fragments are more likely than others to be autonomously stable after, say, limited proteolysis. They have a minimum accessible surface area, which optimizes the contribution of hydrophobicity to their stability, contribution estimated to be 25 cal mol^{-1} Å^{-2} of buried surface area by Chothia (1975), and maximizes the number of internal interactions.

Low accessible surface areas are achieved by small globular proteins, which follow the law (Janin, 1976; Teller, 1976)

$$A_G = 11.1M^{2/3} \qquad (3)$$

relating the accessible surface area A_G to the molecular weight M. The accessible surface area A of nonglobular structures is larger than predicted by eq 3. *The ratio A/A_G is a measure of how globular a structure is.*

We measure the accessible surface area A for all domains defined by interface area scans, assuming that they retain their conformation when the rest of the protein structure is removed. Domains which have A/A_G close to unity are globular as well as independent structures. In cases where A/A_G is significantly larger than 1, a search for globular chain fragments with low values of A/A_G may be performed. An algorithm based on this approach has recently been proposed by Rashin (1981), who gives a different, yet closely related, definition for the globularity index used in the search.

Results

Interface Area Scans and Domain Structures. Interface area scans for several proteins or protein subunits with a variety of domain structures of increasing complexity are presented in Figure 2. Domains bounded by minima of B in the scans are listed on Table I. We give the globularity index A/A_G of each domain and quote the surface areas of domain interfaces. For the larger proteins, connectivity diagrams (Figure 3) illustrate the domain structure as defined by the scans of Figure 2.

BPTI and Cytochrome b_5. The scans of the bovine pancreatic trypsin inhibitor and of cytochrome b_5 are bell shaped and have no significant minimum. These proteins appear as single domains. They are typical small globular proteins (though their shape is far from spherical), with A/A_G near unity.

Immunoglobulin λ Chain. The scan has a minimum at residue 109 where B is very low (250 Å^2) since the variable (1–109) and constant (110–214) domains hardly interact. In

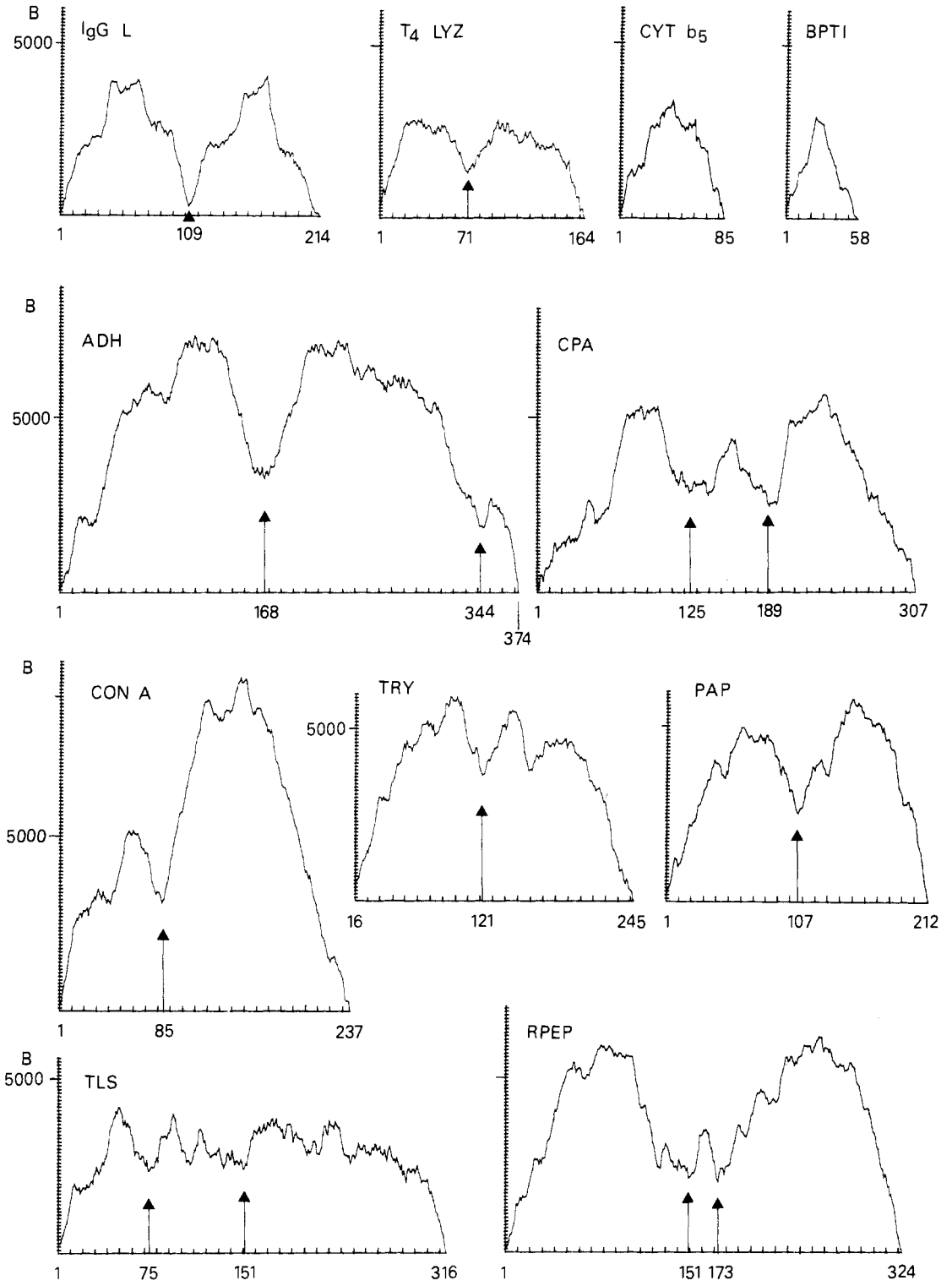

FIGURE 2: Protein interface area scans. Interface areas are calculated from eq 1, with the analytical approximation of Wodak & Janin (1980) and simplified protein models. The values of B in Å^2 are plotted against residue numbers. Minima of B forming domain limits are marked by arrows. The scans represent the λ chain from Fab New (IgG L), phage T4 lysozyme (T_4 LYZ), cytochrome b_5 (CYT B_5), bovine pancreatic trypsin inhibitor (BPTI), alcohol dehydrogenase (ADH), carboxypeptidase A (CPA), concanavalin A (CON A), bovine trypsin (TRY), papain (PAP), thermolysin (TLS), and Rhizopus pepsin (RPEP).

each domain, the scan is similar to that of BPTI or cytochrom b_5. The constant domain is less globular than the variable domain.

Phage T4 Lysozyme and Papain. The division of these proteins into two domains is well established. The scans display it clearly, though the domain interface areas (the value of B at its minimum) are larger than in the λ chain, about 1300 Å^2 in phage lysozyme and 2500 Å^2 in papain. These values are similar to subunit interface areas measured in several protein–protein complexes (Chothia & Janin, 1975; Wodak & Janin, 1980). The domains have A/A_G ratios near unity and form globular structures. Yet, the scans contain significant secondary minima, which may be interpreted as subdomain boundaries. For instance, minima of B at positions 9 and 48 in papain delimit a typical $\beta\alpha\beta'$ unit (Levitt & Chothia, 1976) formed by segment 10–48 of the N-terminal domain.

Trypsin and Serine Proteases. Scans obtained with trypsin, chymotrypsin, elastase, and the bacterial α-lytic protease are

Table I: Domain Structure of Proteins

protein	domains[a]	globularity index A/A_G	interface[b] areas (Å²)
BPTI	1–58	0.99	
cytochrome b_5	1–85	0.96[c]	
immunoglobulin	1–109	1.09	250
λ chain	110–214	1.19	
phage T4	1–71	1.01	1300
lysozyme	72–164	1.04	
papain	1–107	1.02	2500
	108–212	1.04	
trypsin	16–121	1.10	3700
	122–245	1.20	
concanavalin A	1–85	1.17	3200
	86–237	1.19	
carboxy-	1–125	1.04	1950, 2450
peptidase A	126–189	1.04	1350
	190–307	0.96	
thermolysin	1–75	1.01	2350, 50
	76–151	1.00	2400
	152–316	0.98	
Rhizopus pepsin	1–151	1.03	1250, 1200
	152–173	(0.95)[d]	1200
	174–324	1.06	
alcohol	1–168	1.23	1900, 1450
dehydrogenase	169–344	1.15	500
	345–374	(0.90)[d]	
phosphorylase	1–283	1.37	2900, 1950, 1400
	284–483	1.15	250, 1350
	484–658	1.08	4500
	659–841	1.23	
hemoglobin	1–79	1.26	2300
β chain	80–146	1.13	
hen lysozyme	1–38 + 88–129	0.97	1400
	39–87	0.97	

[a] Limits from interface area scans. [b] Interface areas are quoted for pairs of domains. When there are more than two domains, values quoted on a given line correspond to interfaces with each one of the following domains in order. [c] Or 1.08 without the heme group. [d] Values of A/A_G for segments of less than some 40 residues are not meaningful. [e] Limits from Figure 6.

similar. They all have a minimum at residue 121 (chymotrypsin numbering), which defines two domains in serine proteases, often described as made of one β barrel each (Blow, 1971). In trypsin, the two domains 16–121 and 122–245 have globularity indices larger than 1.10. Removal of the N-terminal segment 16–22 from the first domain and of the C-terminal α helix 235–245 from the second brings A/A_G down near unity. These short segments of each domain interact mostly with the other domain.

McLachlan (1979) has shown that the β-barrel unit duplicated in serine proteases may itself be considered as a duplication of a motif formed by four β strands. If the motif forms subdomains, boundaries are expected near residues 58 and 190, where the interface area B is effectively maximum. Indeed, the four-stranded motifs defined by McLachlan (1979) closely interlock to form β barrels, and the interface area scan displays their many interactions. In addition to the minimum at residue 121, the scan has significant minima at residues 83 and 160. They delimit segment 84–121 in the first domain and segment 122–160 in the second, which form large β hairpins with relatively fewer interactions to the rest of the protein structure. The importance of these secondary minima relative to the deeper one at residue 121 decreases from trypsin to elastase to α-lytic protease.

Concanavalin A. A deep minimum of B cleaves the subunit at residue 85. The two domains (Liljas & Rossmann, 1974) are of unequal size; they form β sandwiches of respectively six and seven β strands (Figure 3). The globularity indexes

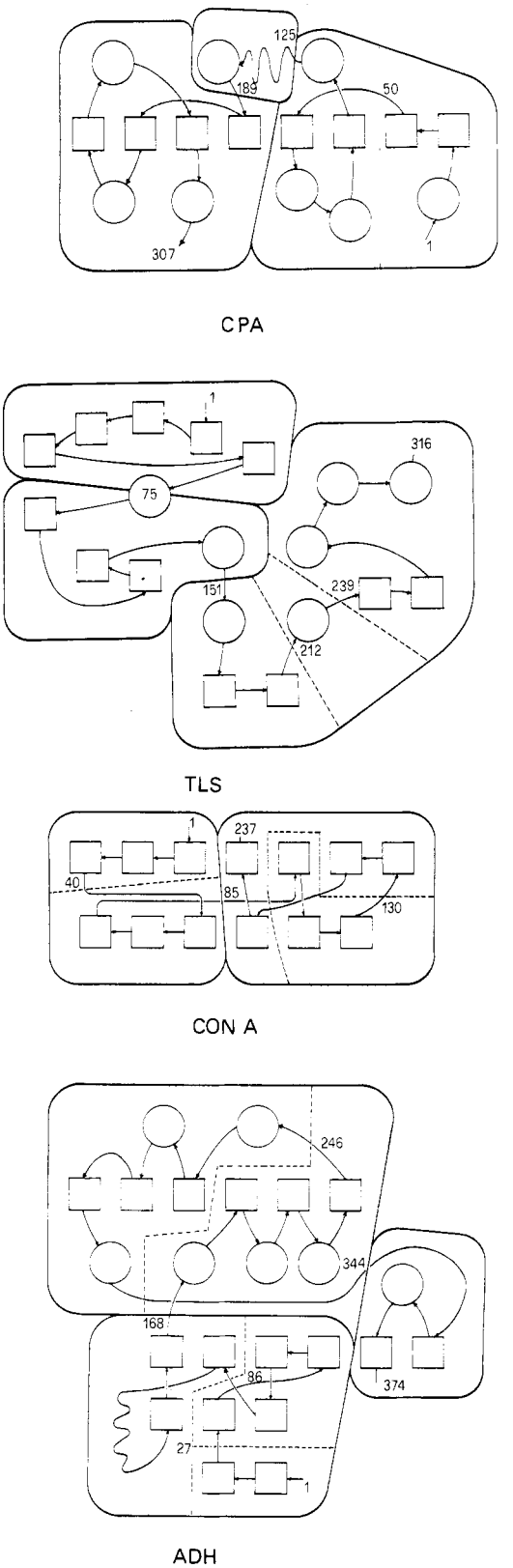

FIGURE 3: Connectivity diagrams. The connectivity diagrams of carboxypeptidase (CPA), concanavalin A (CON A), alcohol dehydrogenase (ADH), and thermolysin (TLS) are inspired from Levitt & Chothia (1976). Squares represent β strands and circles, α helices. Numbers mark limits of domains and subdomains from Figure 2. Envelopes in full line indicate domains and dashed lines, subdomains.

of the domains and that of the whole subunit (1.10) are rather high. In most oligomeric proteins, isolated subunits have larger accessible surface areas than monomeric proteins of the same size (Sprang et al., 1979; Janin, 1979). Secondary minima in the scans at positions 40 and 130 delimit groups of three

β strands in each β sandwich.

Carboxypeptidase A. Two minima of B at positions 125 and 189 define three domains in carboxypeptidase A, in accordance with Liljas & Rossmann (1974). Domains 1–125 and 190–307 represent respectively the right and left halves of the α/β structure (Figure 3), while little secondary structure is found in middle domain 126–189. A secondary minimum isolates segment 1–50 in the first domain; it forms an α helix followed by a β hairpin at the rightmost edge of the large central β sheet. The three domains of carboxypeptidase A are globular, and each domain interacts with the other two, as shown by their large interface areas.

Thermolysin. Thermolysin is described as forming two lobes with about 157 residues in each (Matthews et al., 1974). The interface area scan does have its lowest minimum near the expected position (at residue 151). But, a comparable minimum is seen at position 75 and significant secondary minima elsewhere (positions 109, 212, and 239). The first lobe is therefore composed of two domains. It forms an irregular β sheet folded around helix 66–88. The β sheet may be divided into two parts on each side of the α helix, with the domain boundary at residue 75 occurring in the middle of the helix. This is unusual in our analysis and results from each part of the β-sheet being in contact with one end of helix 66–88. Domains 1–75, 76–151, and 152–306 are globular. The latter also contains a globular subdomain 212–316 (or 240–316) forming a helical bundle.

The area of the interface between the first and third domains is negligible. They do not interact together; both form large interfaces with middle domain 76–151.

Acid Proteases. The interface area scan of Rhizopus pepsin illustrates the domain structure of acid proteases, made of two lobes with half of the polypeptide chain in each lobe. Yet, the scan has two equivalent minima at residues 151 and 173. Only the second is expected on the basis of the description given by Subramanian et al. (1977). Connecting segment 152–173 is a β hairpin placed between the two lobes. It interacts with both, forming equivalent interface areas of about 1200 Å. It corresponds to β-strands q or r of lobe 1 in the nomenclature of Blundell et al. (1979). Domains 1–151 and 174–323 are globular. The globularity index of the short connecting segment is not meaningful.

Andreeva et al. (1979) and Blundell et al. (1979) have noted that the lobes of pepsin have internal symmetry, with each lobe containing twice a structural motif of four β strands. Under the assumption that it forms subdomains, boundaries are expected near residues 60, 150, 235, and 310. The minimum of B at 151 quoted above and a secondary minimum at 65 correspond to the first two boundaries. But, like in serine proteases, the interface areas at the boundaries between motifs in each domain, that is, the values of B near residues 60 and 235, are very large, as the motifs interlock to build the β structure of the domains. Secondary minima of B at residues 126 and 196 are not related to the motifs. They delimit loops (residues 126–151 and 174–196) bulging out of the β structure.

Alcohol Dehydrogenase. The scan has a deep minimum at residue 168, between the catalytic domain 1–168 and the NAD binding domain defined by Rossmann et al. (1975). In their classical description of dehydrogenase subunits, they state that the catalytic domain of alcohol dehydrogenase is discontinuous and contains a C-terminal fragment as well as the first half of the polypeptide chain. The scan does not detect discontinuous domains, but a low minimum of B at residue 344 defines segment 345–374 as a third domain. Its interface area with the NAD binding domain 169–344 is only 500 Å^2, while the interface with domain 1–168 is much larger (1450 Å^2), showing the close association of the C-terminal to the catalytic domain.

Domains 1–168 and 169–344 are not globular, nor is the alcohol dehydrogenase subunit itself (Sprang et al., 1979). The NAD binding domain, which forms most of the subunit–subunit contacts in the dimeric molecule, loses more than its excess accessible surface area in these contacts, but the catalytic domain does not.

Secondary minima of B at positions 27 and 87 delimit β hairpins (residues 1–27 and 28–86), which form subdomains in the catalytic domain. In the NAD binding domain, the interface area scan has a significant minimum at position 246. Cleavage at this point separates the six-stranded α/β structure into two symmetrical parts.

Glycogen Phosphorylase. According to Sprang & Fletterick (1979), the glycogen phosphorylase subunit (841 residues) folds into two structural domains comprising residues 1–489 and 490–841. The first domain can be further divided into two subdomains of about 310 and 160 residues (Weber et al., 1978), with the second forming a glycogen storage site.

Not surprisingly, the interface area scan of this large molecule is complex (Figure 4A). B is minimum at residue 483, in agreement with the crystallographers' description. Yet, the scan has many other significant minima, which may be analyzed by the iterative cleavage algorithm. Interface area scans are run on fragments 1–483 and 484–841; the lowest minima of B in these scans define the limits of subdomains, on which the procedure is repeated. Figure 4A shows scans obtained at the third cycle, together with the primary scan for the whole subunit. Minima of B in these scans define subdomain boundaries. It can be seen that they also correspond to minima of B in the primary scan, though not necessarily to the deepest minima.

The subdomain structure obtained after the third cycle is drawn over the connectivity diagram of the glycogen phosphorylase subunit (Figure 4B). Fragment 1–483 is split first at residue 283, which divides its central nine-stranded β-sheet into five- and four-stranded parts, and defines domain 284–483, equivalent to the glycogen storage domain proposed by Weber et al. (1978). Further cleavage removes the N-terminal end of the chain and places subdomain boundaries at positions 143, 189, and 251 in the first domain and positions 405 and 451 in the glycogen storage domain. The C-terminal fragment 484–841 is arranged around a six-stranded β sheet with the same connectivity as in NAD binding domains of dehydrogenases. The iterative cleavage algorithm splits it at position 658 into two symmetrical parts, as in alcohol dehydrogenase above. Secondary minima of B delimit a bundle of four α helices (positions 484-536) and also the C-terminal segment 811–841, which are not part of the central α/β structure.

Like the glycogen phosphorylase subunit itself (Sprang et al., 1979), domains are not globular. Domain 1–283, which has the highest A/A_G ratio, is heavily involved in the large dimer interface. The domain interface areas quoted in Table I indicate that domain 1–283 also interacts strongly with the three other domains of the subunit. The interface between the two halves 484–658 and 659–841 of the C-terminal fragment is particularly extensive (4500 Å^2), which explains why they have been considered as a single very large domain.

Domains and Exons in Globins and in Hen Lysozyme. Globins. Figure 5 shows the interface area scan of the β chain of human hemoglobin. It has a well-defined minimum near residue 79. Scans of the hemoglobin α chain of myoglobin

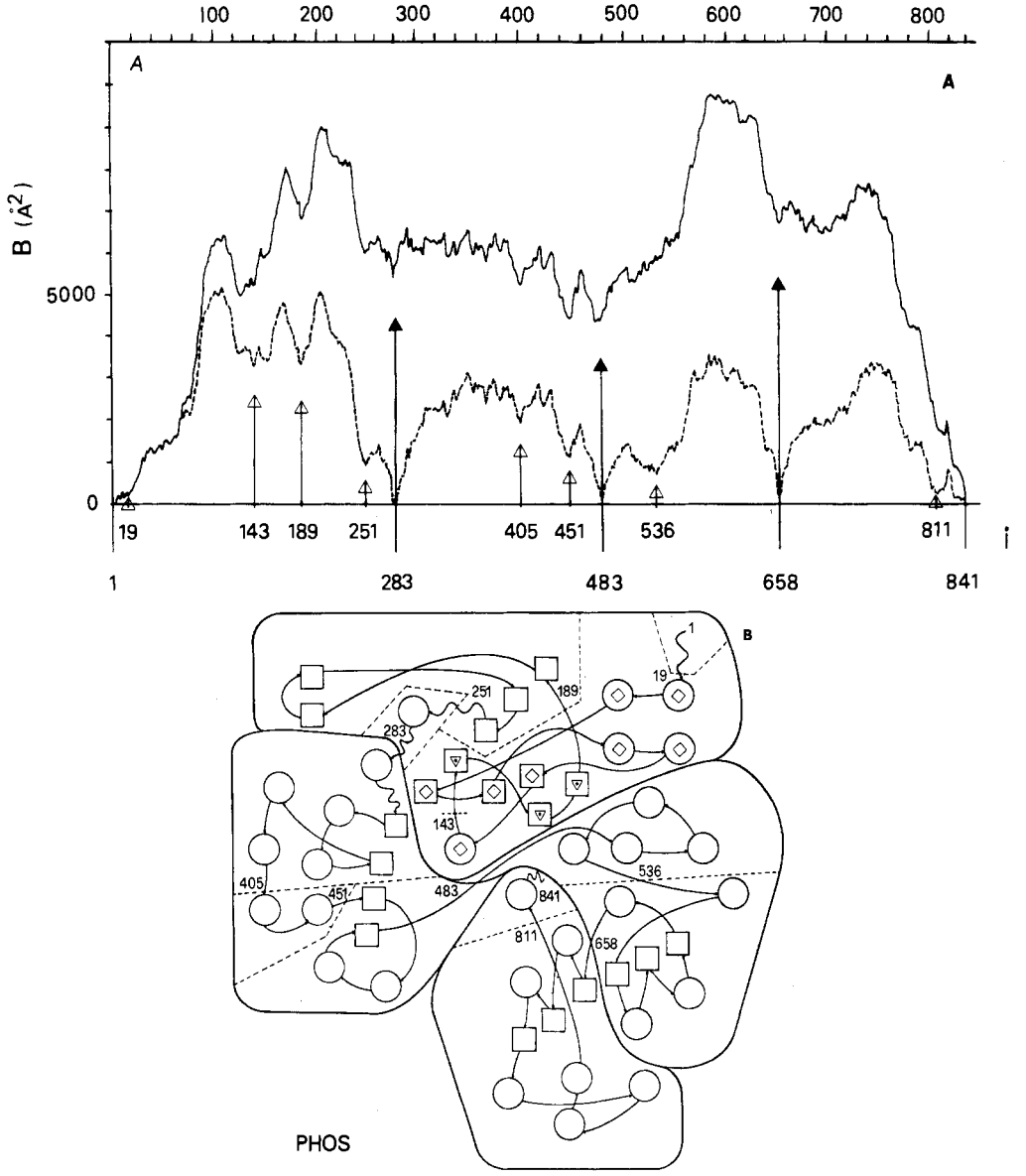

FIGURE 4: Glycogen phosphorylase. The interface area scan of the whole subunit is shown in full line on (A), and scans of fragments are dashed. Full arrows mark limits of domains and empty arrows, limits of subdomains. The connectivity diagram (B) is adapted from Sprang & Fletterick (1979). Subdomains 20–143 and 144–189 cannot be easily represented by an envelope on this diagram. Helices and β strands in these subdomains are labeled with diamonds and inverted triangles.

and of leghemoglobin are very similar to the one shown. The minimum is at position 74 in the α chain (equivalent to position 79 in Hbβ) and at position 85 in the other two globins. Thus, the scans divide the globin fold into two domains, at the junction of the E and F helices, with the heme group at the domain interface—as originally proposed by Wetlaufer (1973) and by Ptitsyn & Rashin (1975). The interface area is about 2300 Å^2, not counting the heme group, a value similar to the domain interface area of papain or thermolysin.

The interface area scans define domains in globins just as clearly as they do in the two proteins just cited. Yet, the β-chain domains are less globular than those of papain: they have fairly large A/A_G ratios. We searched for globular chain segments with a minimum A/A_G ratio and found one in the first domain, segment 19–69 of Hbβ (and similar segments in other globins), which includes helices B, C, and D and the part of helix E that packs onto helix B. This segment has A/A_G close to unity and forms a globular core within the first domain. In the second domain, segment 107–138, a hairpin formed by G and H plays the same role. Similar results have been obtained by Rashin (1981), while Gō (1981) draws different conclusions from diagonal plots.

Rossmann & Argos (1975) observe a structural homology between α-helical parts in cytochrom b_5 and globins. They define a heme binding unit which in the hemoglobin β chain, comprises about 48 residues between positions 12 and 107. This part of the structure forms both sides of the heme binding pocket, of which each of the structural domains 1–79 and 80–146 forms one side. There is therefore no correlation between the heme binding unit defined by Rossmann & Argos and the structural domain found here. It has been proposed (Blake, 1979; Eaton, 1980) that the heme binding unit corresponds to the central exon of globin genes. This exon codes for segment 31–104 in several Hbβ genes. Positions 31 and 104 are near maxima in the interface area scan of Figure 5. With the heme group removed, segment 31–104 has an open structure with a high globularity index A/A_G of 1.4. It is hardly more globular with the heme present (A/A_G = 1.3). Prepared by limited proteolysis of the β chain, segment 31–104 has been observed to bind heme (Craik et al., 1980). However, the complex cannot be made to bind oxygen, and its circular dichroism indicates little α-helical structure, unless the other

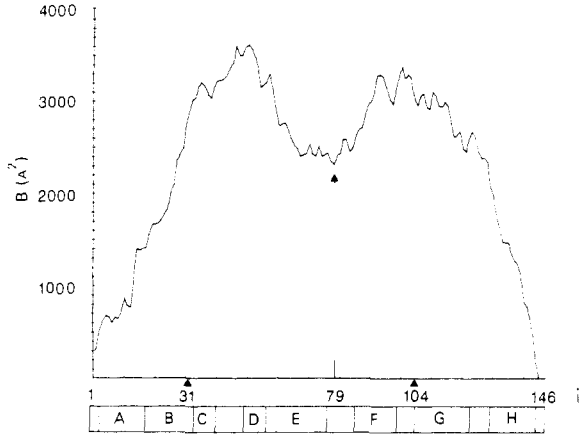

FIGURE 5: Hemoglobin β chain. The interface area scan of the human deoxy β chain shown here is obtained with the heme group omitted from the surface area calculations. A scan made with the heme group attached to His-92 would be very similar; the minimum of B near position 79 would be less pronounced due to interactions between the heme and the two halves of the protein, which increase B by about 300 Å^2. Positions 31 and 104 mark the limits of the central exon in several Hbβ genes. Helical segments are indicated below the horizontal axis.

two exon products and the α chain are added (Craik et al., 1981). Thus, the tertiary structure of the central exon product, if it has one, is unlikely to resemble the native one.

Hen Lysozyme. Hen lysozyme has been described as forming two domains, a continuous central domain 39–87 and a discontinuous domain comprising the N- and C-terminal segments (Phillips, 1967; Wetlaufer, 1973). The interface area scan has a significant minimum at residue 38, but not at residue 87, and its general appearance is that of a single domain protein. Yet, the central domain is well individualized in the structure. It shows as a minimum of B in a two-dimensional search, where a segment of variable length L is moved along the polypeptide chain (Figure 6), a technique applicable to proteins containing discontinuous domains interrupted by continuous ones. The low A/A_G ratio confirms the globular structure taken by segment 39–87, which forms a three-stranded β sheet with an α helix packed onto it.

Rossmann & Argos (1976) find structural homology between a region of hen lysozyme that contains most of the residues binding the substrate, and the first domain of phage T4 lysozyme. In their comparison, segment 25–101 of the hen enzyme corresponds to 1–73 of the phage enzyme. The substrate binding unit is not equivalent to the central domain 39–87, which forms only one side of the binding site. The situation is the same as in globins, and here again, a correlation between ligand binding regions of the protein molecule and exons in the gene has been proposed (Jung et al., 1980). Introns in the hen lysozyme gene occur at positions 28, 81, and 108. Three of the four exons code for chain segments of less than 30 residues. Their extremities fall in the middle of α helices. Segment 28–108 coded for by the second and third exons could conceivably form a stable structure in the absence of the rest of the chain and be identified to the substrate binding unit 25–101 (Matthews et al., 1981). Yet, neither this segment nor any of the exon products show as a structural domain in our analysis. They have large interfaces with the rest of the lysozyme structure (Figure 6) and do not appear as globular autonomous regions of the protein.

Discussion

In all proteins studied here, interface area scans define essentially the same domains as does inspection of physical models of the protein structure, at least for continuous domains. A comparison of Table I with published descriptions of protein structures reveals a difference only in thermolysin, where we show that the N-terminal half forms two distinct globular units with a small interface, while it has been described as a single domain. In alcohol dehydrogenase, the discontinuous catalytic domain is split into two, but the domain interface areas show the association of the N- and C-terminal fragments that form the catalytic domain.

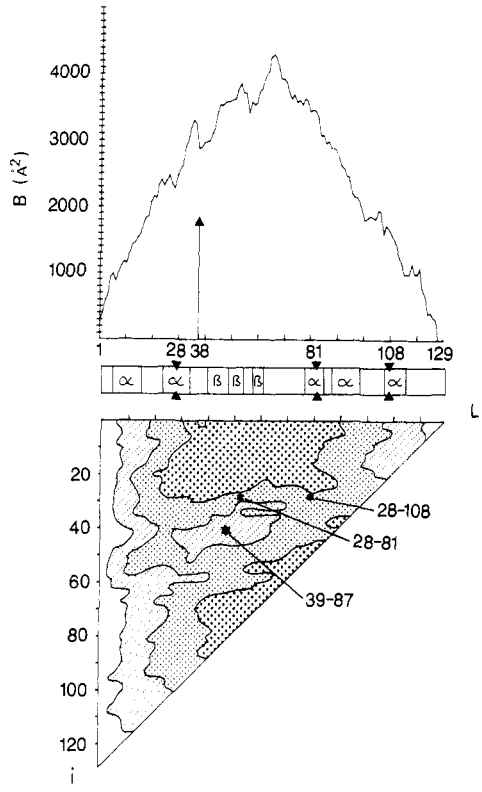

FIGURE 6: Hen lysozyme. The top part of the figure is an interface area scan obtained as in Figure 2. It corresponds to $i = 1$ in the two-dimensional map on the bottom, where the interface area B of a chain segment with the rest of the protein is plotted as a function of the N-terminus position i and of the length L of the segment. B is contoured at 1000, 2000, and 3000 Å^2, with shadings of increasing intensity. A large star marks the minimum of B corresponding to domain 39–87. Small stars correspond to exon products. Exon limits (Jung et al., 1980) and secondary structure in hen lysozyme are indicated in between the two parts of the figure.

Visual inspection relies on a subjective estimate of the extent to which parts of the protein molecule interact. *Interface area scans give a quantitative measurement of the same.* They provide a simple and convenient way of cleaving the protein structure into continuous domains. They may be used iteratively to yield more details than can be seen easily by inspection, as we show for glycogen phosphorylase. In addition, the procedure may be extended when necessary to detect continuous domains in the presence of discontinuous ones, for instance, in hen lysozyme.

Surface area is a geometric concept. A geometrical definition of domains may not coincide with biochemical or functional definitions. Yet, a correlation established previously between accessible surface areas and hydrophobic free energies (Hermann, 1972; Chothia, 1974) provides a link between the two. Interface area scans identify fragments with a minimum hydrophobic contribution to their association. Conformational energy calculations could be used to calculate the contributions of van der Waals and electrostatic interactions. For most purposes, however, interface areas, which are correlated to the number of contacts (Wodak & Janin, 1978) and are much less sensitive to errors in atomic positions, may be taken as crude

estimates of van der Waals interactions. Thus, algorithms using surface area criteria have been shown to succeed in finding optimal contacts between protein subunits (Wodak & Janin, 1978) or between β sheets and α helices in tertiary structure prediction (Richmond & Richards, 1978; Cohen et al., 1980). In these studies, structural elements with a maximum of interaction are considered, while we search here for chain segments with a minimum of interaction, but otherwise the approach is the same.

Algorithms for the location of domains starting from atomic coordinates have been proposed by Crippen (1978) and by Rose (1979). Crippen applies a clustering algorithm to short-chain segments. The metrics used in the clustering rely on the definition of a "packing density" derived from distances between α-carbon atoms. The algorithm generates an ascending hierarchy of associations between segments, clusters of which may be assimilated to domains or subdomains. Such domains are almost unavoidably discontinuous, which makes a comparison with our results difficult. Rose uses an algorithm of binary divisions of the polypeptide chain, which creates a descending hierarchy of continuous segments. His approach is therefore closer to ours. Yet, the definition of cleavage points in Rose's procedure relies on a very different technique. A plane parallel to one of the principle axes of inertia is chosen so as to cleave the three-dimensional structure into two parts. A bias is introduced to generate fragments of about equal sizes. A comparison of the data in Table I of Rose (1979) with the results presented here shows that the first cleavage occurs at or near a minimum of B in the interface area scans, for a number of proteins. The agreement is good in the sense that the higher levels of Rose's hierarchies coincide with domains or subdomains defined by interface area scans. But our algorithm stops at points where the scans contain no significant minima of B, while Rose's may cleave any chain segment. It cleaves structures that we find to form single domains, and in these cases, it cleaves at a maximum of B instead of a minimum. BPTI is cleaved at residue 27, near the maximum of the interface area (Figure 2). In other proteins, cleavage occurs near minima of B at the higher levels of the descending hierarchy but near maxima of B at lower levels.

Like us, Crippen (1978) and Rose (1979) implement Wetlaufer's definition of structural domains and aim to find globular autonomous units in protein structures. Rossmann and his collaborators (Rao & Rossmann, 1973; Rossmann & Argos, 1975, 1976) have taken a different approach. They search for structural homologies between parts of proteins and identify structures that are repeated from one protein to another, or within a protein, and may be associated with a specific function such as ligand binding. Superposition algorithms (Rossmann & Argos, 1976; Remington & Matthews, 1978; MacLachlan, 1979) are designed to detect such repeated structures. We have already discussed several cases of this sort: the heme binding unit in globins and cytochrom b_5, the substrate binding unit in hen and phage T_4 lysozymes, and the duplicated structural units within the domains of trypsin and of acid proteases. We have shown that there is little, if any, correlation between these repeated units and the structural domains or subdomains that we define by surface area criteria. Repeated units are often not globular, and they form large interfaces with the rest of the protein structure. In the absence of detectable homologies in the amino acid sequence, their presence may result from convergent evolution as well as from gene duplication.

The discovery of separate DNA sequences coding for each immunoglobulin domain has led to the hypothesis that coding sequences (exons) correspond to domains in eukaryotic proteins (Gilbert, 1978). Introns are found within domains as well as between in the immunoglobulin genes. Reshuffling of the coding sequences through somatic rearrangement of the genome is important in antibody formation. A generalization of these mechanisms to protein evolution meets with many difficulties, not the least, the absence of introns in the genome of prokaryotes. Identical domain structures are observed in bacterial and animal proteins, for instance, in serine proteases or in glyceraldehyde-3-phosphate dehydrogenases (Biesecker et al., 1977), which suggests that introns have had no effect on their evolution after prokaryotes diverged from eukaryotes. For hemoglobin and for hen lysozyme, there is no correlation between exons and the structural domains defined here. The significance of the correlation observed between exons and heme or substrate binding units remains to be assessed. In any case, exon products in globins and in hen lysozyme do not constitute domains in the sense of Wetlaufer (1973).

Underlying Wetlaufer's definition of domains as globular autonomous regions of proteins is a model of protein folding for which evidence has accumulated. Parts of the polypeptide chain form stable tertiary structures, which are intermediates in the formation of the native structure, at least in large proteins (Kirschner & Bisswanger, 1976; Baldwin & Creighton, 1980). Structural requirements for such intermediates are similar to those governing the stability of small globular proteins, hence the search for chain segments making a maximum of internal interactions and a minimum of external ones. In contrast, the search for structural homology within or between proteins has an evolutionary background: structural homology suggests divergence of domains from a common ancestral protein. The hypothesis is especially strong when homologous structures perform homologous functions (Rossmann et al., 1974).

Let us reconcile the two approaches in a sort of molecular equivalent to the classical Haeckel theory of ontogeny: domains, which *are intermediates on the path leading from simple to complex* during protein folding, have played the same role during protein evolution through gene fusion. That we find no systematic coincidence between repeating structural units and globular autonomous regions may add significance to the cases where the coincidence exists: the domains of serine proteases and of acid proteases, the NAD binding domains of dehydrogenases (also found in phosphorylase) and the immunoglobulins.

Acknowledgments

This work was initiated at a workshop on protein folding held at the Centre Européen de Calcul Atomique et Moléculaire, Orsay, France, 1979. We acknowledge stimulating discussions with A. A. Rashin, G. Rose, and C. Sander.

References

Andreeva, N. S., & Gutschina, A. E. (1979) *Biochem. Biophys. Res. Commun. 87*, 32–42.

Baldwin, R. L., & Creighton, T. E. (1980) in *Protein Folding* (Jaenicke, R., Ed.) pp 217–261, Elsevier, Amsterdam.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Biesecker, G., Harris, J. I., Thierry, J. C., Walker, J. E., & Wonacott, A. J. (1977) *Nature (London) 266*, 328–333.

Blake, C. C. F. (1979) *Nature (London) 277*, 598.

Blow, D. (1971) *Enzymes, 3rd Ed. 3*, 185–212.

Blundell, T. L., Sewell, B. T., & McLachlan, A. D. (1979) *Biochim. Biophys. Acta 580*, 24–31.
Chothia, C. (1974) *Nature (London) 248*, 338–339.
Chothia, C. (1975) *Nature (London) 254*, 304–305.
Chothia, C., & Janin, J. (1975) *Nature (London) 256*, 705–708.
Cohen, F. E., Sternberg, M. J. E., & Taylor, W. (1980) *Nature (London) 285*, 378–382.
Craik, C. S., Buchman, S. R., & Beychok, S. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 1384–1388.
Craik, C. S., Buchman, S. R., & Beychok, S. (1981) *Nature (London) 291*, 87–90.
Crippen, G. M. (1978) *J. Mol. Biol. 126*, 315–332.
Eaton, W. A. (1980) *Nature (London) 284*, 183–185.
Edelman, G. M., Cunningham, B. A., Gall, W. E., Gottlieb, P. D., Rutishauser, U., & Waxdal, M. J. (1969) *Proc. Natl. Acad. Sci. U.S.A. 63*, 78–85.
Gilbert, W. (1978) *Nature (London) 281*, 501.
Gō, M. (1981) *Nature (London) 291*, 90–93.
Hermann, R. B. (1972) *J. Phys. Chem. 76*, 2754–2759.
Janin, J. (1976) *J. Mol. Biol. 105*, 13–14.
Janin, J. (1979) *Bull. Inst. Pasteur (Paris) 77*, 337–374.
Jung, A., Sippel, A. E., Grez, M., & Schütz, G. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 5759–5763.
Kirschner, K., & Bisswanger, H. (1976) *Annu. Rev. Biochem. 45*, 143–166.
Lee, B. K., & Richards, F. M. (1971) *J. Mol. Biol. 55*, 379–400.
Levitt, M. (1976) *J. Mol. Biol. 104*, 59–107.
Levitt, M., & Chothia, C. (1976) *Nature (London) 261*, 552–557.
Liljas, A., & Rossmann, M. G. (1974) *Annu. Rev. Biochem. 43*, 475–507.
Matthews, B. W., Weaver, L. H., & Kester, W. R. (1974) *J. Biol. Chem. 249*, 8030–8044.
Matthews, B. W., Grütter, M. G., Anderson, W. F., & Remington, S. J. (1981) *Nature (London) 290*, 334–335.
McLachlan, A. D. (1979) *J. Mol. Biol. 128*, 49–79.
Phillips, D. C. (1967) *Proc. Natl. Acad. Sci. U.S.A. 57*, 484–495.
Ptitsyn, O. B., & Rashin, A. A. (1975) *Biophys. Chem. 3*, 1–20.
Rao, S. T., & Rossmann, M. G. (1973) *J. Mol. Biol. 76*, 241–256.
Rashin, A. A. (1981) *Nature (London) 291*, 85–87.
Remington, S. J., & Matthews, B. W. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 2180–2184.
Richmond, T. J., & Richards, F. M. (1978) *J. Mol. Biol. 119*, 537–555.
Rose, G. D. (1979) *J. Mol. Biol. 134*, 447–470.
Rossmann, M. G., & Liljas, A. (1974) *J. Mol. Biol. 85*, 177–181.
Rossmann, M. G., & Argos, P. (1975) *J. Biol. Chem. 250*, 7525–7532.
Rossmann, M. G., & Argos, P. (1976) *J. Mol. Biol. 105*, 75–95.
Rossmann, M. G., Moras, D., & Olsen, K. W. (1974) *Nature (London) 250*, 194–199.
Rossmann, M. G., Liljas, A., Bränden, C. I., & Banaszak, L. J. (1975) *Enzymes, 3rd Ed. 11*, 61–102.
Sprang, S., & Fletterick, R. J. (1979) *J. Mol. Biol. 131*, 523–551.
Sprang, S., Yang, D., & Fletterick, R. J. (1979) *Nature (London) 280*, 333–335.
Subramanian, E., Swan, I. D. A., Liu, M., Davies, D. R., Jenkins, J. A., Tickle, I. J., & Blundell, T. L. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 556–561.
Teller, D. C. (1976) *Nature (London) 260*, 729–731.
Weber, I. T., Johnson, L. N., Wilson, K. S., Keats, D. G. R., Wild, D. L., & Jenkins, J. A. (1978) *Nature (London) 274*, 433–437.
Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697–701.
Wodak, S. J., & Janin, J. (1978) *J. Mol. Biol. 124*, 323–342.
Wodak, S. J., & Janin, J. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 1736–1740.