Critical Reviews in Biochemistry and Molecular Biology, 31(2):127–152 (1996)

Protein-Protein Interfaces: Architectures and Interactions in Protein-Protein Interfaces and in Protein Cores. Their Similarities and Differences

Chung-Jung Tsai,[1] Shuo Liang Lin,[1] Haim J. Wolfson,[2] and Ruth Nussinov[3,4]

[1]Laboratory of Mathematical Biology, NCI-FCRDC, Bldg. 469, Room 151, Frederick, MD 21702; [2]Computer Science Department, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel; [3]Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv 69978, Israel; [4]Laboratory of Mathematical Biology, SAIC, NCI-FCRDC, Bldg. 469, Room 151, Frederick, MD 21702

Referee: Dr. Mike Pitman, IBM, T.J. Watson Research Center, Yorktown Heights, New York

* Correspondence should be addressed to **R. Nussinov** at NCI-FCRDC, Bldg. 469, Room 151, Frederick, MD 21702.

ABSTRACT: Protein structures generally consist of favorable folding motifs formed by specific arrangements of secondary structure elements. Similar architectures can be adopted by different amino acids sequences, although the details of the structures vary. It has long been known that despite the sequence variability, there is a striking preferential conservation of the hydrophobic character of the amino acids at the buried positions of these folding motifs. Differences in the sizes of the side-chains are accommodated by movements of the secondary structure elements with respect to each other, leading to compact packing. Scanning protein-protein interfaces reveals that similar architectures are also observed at and around their interacting surfaces, with preservation of the hydrophobic character, although not to the same extent.

The general forces that determine the origin of the native structures of proteins have been investigated intensively. The major non-bonded forces operating on a protein chain as it folds into a three-dimensional structure are likely to be packing, the hydrophobic effect, and electrostatic interactions. While the substantial hydrophobic forces lead to a compact conformation, they are also nonspecific and cannot serve as a guide to a conformationally unique structure. For the general folding problem, it thus appears that packing is a prime candidate for determining a particular fold. Specific hydrogen-bonding patterns and salt-bridges have also been proposed to play a role. Inspection of protein-protein interfaces reveals that the hallmarks governing single chain protein structures also determine their interactions, suggesting that similar principles underlie protein folding and protein-protein associations. This review focuses on some aspects of protein-protein interfaces, particularly on the architectures and their interactions. These are compared with those present in protein monomers. This task is facilitated by the recently compiled, non-redundant structural dataset of protein-protein interfaces derived from the crystallographic database. In particular, although current view holds that protein-protein interfaces and interactions are similar to those found in the conformations of single-chain proteins, this review brings forth the differences as well. Not only is it logical that such differences would exist, it is these differences that further illuminate protein folding on the one hand and protein-protein recognition on the other. These are also particularly important in considering inhibitor (ligand) design.

KEY WORDS: protein-protein interfaces, protein architectures, structural motifs, protein-protein interactions, interface-dataset.

1040-9238/96/$.50

I. INTRODUCTION

Over the last few years, the number of studies of protein-protein interfaces has increased very rapidly (e.g., see Janin et al., 1988; Argos, 1988; Neet and Timm; 1994; Jackson and Sternberg, 1995; Wallqvist et al., 1995). There are several reasons for focusing on protein-protein interfaces (Janin and Chothia, 1990). First, one would like to comprehend the nature of protein-protein recognition and the principles that underlie molecular associations. Understanding these is also likely to have practical implications, such as in docking and in molecular design (e.g., Connolly, 1986; Goodsell and Olson, 1990; Cherfils and Janin, 1993; Duncan and Olson, 1993a,b; Norel et al., 1994; 1995; Fischer et al., 1995a). Second, one would like to be able to investigate the similarities and the differences between protein folding and molecular recognition. There are numerous indications that the two processes are not dissimilar (e.g., Walls and Sternberg, 1992; Young et al., 1994). Figuring these out may be forecast to aid in the prediction of molecular associations, and, at the same time, afford an added insight into protein folding.

Most investigations of protein-protein interfaces have been carried out on either specific interfaces, whether between protein subunits, or between receptor-ligand (inhibitor) associations (e.g., Derrick and Wigley, 1992; Milburn et al., 1993; Lodi et al., 1994; Clore et al., 1994), or on relatively small collections of interfaces (e.g., Argos, 1988; Janin et al., 1988; Miller, 1989; Lin et al., 1995; Wallqvist et al., 1995). A major difficulty in carrying out comprehensive studies on protein-protein interfaces resides in obtaining a non-redundant structural dataset of interfaces (Tsai et al., 1996a). As the crystallographic database contains many similar, or practically nearly identical, structures, the first step in carrying out any statistically based study, is to execute massive comparisons between the interfaces, retaining a representative dataset. Such an approach has already been adopted a number of years ago for obtaining a non-redundant dataset of single-chain (monomer) proteins (Boberg et al., 1992; Pascarella and Argos; 1992; Holm et al., 1993; Orengo et al., 1993; Orengo, 1994; Fischer et al., 1995b). However, the problem of constructing such a non-redundant dataset for interfaces is much more complex. This review describes the derivation of such a structural dataset, focusing on searches for recurring architectural motifs in protein-protein interfaces and their comparisons with those found in protein cores. While there are numerous studies on the types of motifs present in cores (see references cited above), investigations as to their presence, or absence, in protein-protein interfaces have not been carried out. Yet, such a research can be very rewarding. From the similarities, and dissimilarities, between these not only can we abstract useful information for molecular docking and design, but as here more than one chain is involved in the formation of the motif, footprints of the folding process may be obtained as well (Lin et al., 1995). Here we discuss such similarities and dissimilarities in the architectural motifs and in the types of interactions between interfaces and cores and explore how to abstract and how to potentially utilize this information both toward recognition and toward protein folding.

Protein structures follow some general principles (Dill, 1990; Rose and Wolfenden, 1993). The interior of proteins is hydrophobic, forming a hydrophobic core. The hydrophilic backbones are hydrogen-bonded, forming secondary structure elements. The hydrophobic side-chains of the α-helices and of the β-sheets frequently interact to form higher-order secondary structure motifs. The actual amino acid sequences that

fold into these motifs differ, providing some evidence for convergent evolution. Examination of protein structures has indicated that the number of motifs is limited, falling into some distinct classes (Chothia, 1992; Pascarella and Argos, 1992; Rufino and Blundell, 1994; Alexandrov and Go, 1994; Crippen and Maiorov, 1995). Recurrence of particular arrangements of secondary structure elements suggests that they are highly favorable (Finkelstein and Ptitsyn, 1987). Loops connecting the secondary structure elements are often rich in charged, and polar, hydrophilic amino acids. They are generally found on the surface, exposed to the solvent. Inspection of protein-protein interfaces demonstrates that motifs observed in the interior of proteins are present at their interfaces, where one part of the motif is contributed by one protein chain, and the second part by the other. Hydrophobic interactions are observed in these, protein-protein interface architectures, in a manner analogous to that observed in the interior of the proteins (Tsai et al., 1996b). The hydrophobic interactions, shown to play a critical role in the stabilization of these secondary structure architectures at the interfaces, originate from the surfaces of both molecules. Loops, which are located on protein surfaces, tend to be accommodated at the edges of the interacting protein interfaces, rather than in their midst. Nevertheless, we have recently shown (Tsai et al., 1996b) that while the hydrophobic interactions are important for protein-protein recognition, they are not as dominant as for protein folding. On the other hand, hydrogen bonds and ion pairs play a more important role in binding than in folding (see also Horton and Lewis, 1992).

Secondary structure motifs have been noted at protein-protein interfaces. There are several well-known examples. Rop consists of two polypeptide chains (Banner et al., 1987). A four-helix bundle comprises their interface, with two α-helices originating from each chain. The constant domains of the immunoglobulins form an antiparallel compressed β-barrel at their interface (Novotny et al., 1983). A barrel is observed between the variable domains of the heavy and light chains. Here we show that these are not isolated incidents, but rather, a general rule: examples include four-helix bundles, α/β structures, β-barrels, interlaced β-sheets, with the alternating β-strands originating from the different protein chains and extensions of β-sheets at chain-chain interfaces. The hydrophobic patches on protein surfaces are involved in the interaction between the secondary structure elements in these architectural motifs.

In particular, this review focuses on a detailed description of one motif, the four-helix bundle, which has been well characterized both in protein monomers (Presnell and Cohen, 1989; Harris et al., 1994; Lin et al., 1995) and in protein-protein interfaces (Lin et al., 1995). While, in general, the architectures are quite similar, differences have been observed as well.

II. THE CRITERIA FOR INTERFACE SELECTION

Several criteria were separately explored to define interacting residues at the interface between two protein chains. When (1) the distance between two C_α atoms, or (2) between two C_β ones, one from each of the chains, is less than 9.0 Å, the two corresponding residues are flagged as interacting residues. (3) Alternatively, if the distance between any two atoms belonging to two residues from the different chains is less than 5.0 Å, or (4) less than the sum of their corresponding van der Waals radii plus 0.5 Å, the residues are considered as belonging to the interface. (5) If the van der Waals energy between the residues is less than –0.5 kcal/mol, they are categorized as belonging to the interface. (The van der Waals

parameterization employed throughout this work is taken from CHARMM, Brooks et al., 1983.) Regardless of the criteria adopted, the set of selected interface residues is quite consistent. As the fourth criterion is the most indicative in terms of residue-residue contact, it has been followed in this work. In order to be able to analyze the types of architectures at the interfaces, and to compare them with those found in the interior of the proteins, large enough protein scaffolds are needed. For this reason, we label any residue other than the interfaced ones as a "nearby" residue if the distance between its C_α, and a C_α of an interfaced residue is less than 6.0 Å. For the classification of the architectures at the interfaces, both the interacting amino acids and the nearby ones are considered (dark colors in Plate 1).*

III. PROTEIN-PROTEIN INTERFACES AND PROTEIN CORES: SIMILAR ARCHITECTURES

This interface-picking procedure has been applied to all entries in the Brookhaven Protein Data Bank (Bernstein et al., 1977). Interfaces consisting of a relatively large number of residues have been inspected. Plates 1a to h demonstrate some examples of the types of motifs at two-chain interfaces. A right-twisted four-helix bundle is observed between the two chains of the Fis DNA binding protein (PDB code 1 fia, Kostrewa et al., 1992; Plate 1A). An antiparallel β-sheet with two α-helices on one side and a third α-helix on the other side is shown at the interface of two chains, D and E, of an enterotoxin (1 lta, Sixma et al., 1993; Plate 1B). Such a motif is frequently encountered in many proteins. Here part of the sheet is contributed by one chain, and its continuation by the second chain. The β-strands at the interface are hydrogen bonded to each other. Plate 1C displays a β-barrel composed of six antiparallel β-strands at the interface of the two chains of ubiquitin (1 aar, Cook et al., 1992). Three strands are contributed by each subunit. A compressed barrel architecture is observed at the interface of the constant domains of the immunoglobulins, C_L and C_{H1} (1 dba, Arevalo et al., 1993), although here four β-strands are contributed by each domain. The two β-sheets are rotated with respect to each other by 45° (Plate 1D). This motif occurs frequently in protein cores (e.g., in retinol binding proteins) (Newcomer et al., 1984). A barrel topology is observed at the interface of the variable domains of the heavy and light chains in the immunoglobulins. Here the orientation between the two half barrels, contributed by the two chains, is 60° (1 dba, Arevalo et al., 1993; Plate 1E). Two examples of interlaced β-sheets, where alternating strands are contributed by the two interacting chains at their interface, are displayed in Plates 1F and 1G. Plate 1F depicts interlaced β-sheets between the α and β chains of the first monomer of the pea lectin (2 ltn, Prasthofer et al., 1989). Significantly shorter interlaced strands are observed in an HIV type 2 protease (1 ivp, Mulichak et al., 1993; Plate 1G). β-sheets are an extremely common motif found in numerous protein structures. Plate 1H displays a β-sandwich architecture (2 pab, Blake et al., 1978). Hydrogen bonds between the two chains are observed between the lower parts of the backbones. The interior of this β-sandwich contains considerably more hydrophilic amino acids than the other motifs noted above.

IV. HYDROPHOBIC INTERACTIONS AT THE INTERFACES

The hydrophobic interactions at the interfaces between these subunits are high-

* Plate 1 appears after page 134.

lighted in Plates 2A to H.* These interacting hydrophobic patches contribute directly to the stabilization of the motifs. The strictly hydrophobic core of the four-helix bundle, arising from the two chains, is seen in Plate 2A. The amphipathic α-helices of the enterotoxin are seen in Plate 2B. The hydrophobic residues of the helices point toward the β-sheet. A hydrophobic core exists inside the β-barrel at the interface of the two chains of ubiquitin (Plate 2C). Interacting hydrophobic residues are observed between the compressed β-barrel halves of the constant domains of immunoglobulins, with two phenylalanines contributed by each (Plate 2D). The hydrophobic core of the β-barrel-like architecture of the variable domain of the immunoglobulins is clearly seen in Plate 2E, and is manifested in Figures 1 and 2 (discussed below). The hydrophobic side-chains of the interlaced β-strands at the interface of the two chains (Plate 2F,G), and between the sheets of the β-sandwich in the lectin (Plate 2F), point toward each other. This mode of packing, burying the hydrophobic side-chains between interacting α-helices, between β-sheets, inside barrels, and between α-helices and β-strands has been observed frequently in protein cores. β-strands have also been shown to twist to maximize their hydrophilic interactions (Finkelstein and Ptitsyn, 1987).

Figures 1 and 2 provide some measure of quantification of two parameters of these hydrophobic interactions for these interfaces. The observed vs. expected frequency of interaction between hydrophobic residues at the interface is given in Figure 1. The amino acids have been classified as hydrophobic (Ala, Val, Leu, Ile, Phe, Trp, Met, and Pro), designated H; hydrophilic (Arg, Lys, Glu, Gln, Asp, Asn, Ser, Thr, His), labeled P; and amphiphilic (Gly, Tyr, and Cys), designated A (Cornette et al., 1987).

* Plates 2A to 2H appear following page 134.

Interacting amino acids are those satisfying condition (4) noted above. In the calculation of the expected values, we have made two assumptions. First, interfacing residues can be only those residing on protein surfaces. Second, residues on the surface are randomly distributed. With these two assumptions, the probability of hydrophobic interactions at the interface can be calculated as $(H_a + A_a) \cdot (H_b + A_b)/(H_a + A_a + P_a) \cdot (H_b + A_b + P_b)$, where H, P, and A are the population of each group of residues on the surfaces of chains a and b. The observed hydrophobic interactions are the sum of interacting hydrophobic and/or amphiphilic residues divided by the total number of interacting residues across the two chain interface. Inspection of Figure 1 indicates that while the magnitudes of the hydrophobic interactions vary, the observed interactions are consistently more frequent than those expected by chance.

Figure 2 is a histogram displaying the percentage of hydrophobic residues on the molecular surfaces that have been buried in the interface after association of the two chains and complex formation. To obtain the histogram, the crystal structures of these protein chains have been scanned, and their residues classified as accessible to the solvent or as buried in the interior. The criterion for such a classification is their solvent accessible areas. The residue accessibility is defined as the ratio of its surface accessible area (ASA) in the crystal structure to the ASA the residue would have in an extended polypeptide chain. The surface accessibility definition of Lee and Richards (1971) has been adopted here, and the accessible areas have been calculated using the Shrake and Rupley (1973) algorithm, as implemented by Miller et al. (1987). A residue was classified as residing on the surface if its ASA in the structure is above 20% of its reference Gly-X-Gly ASA in the extended conformation (Lodi et al., 1994). If

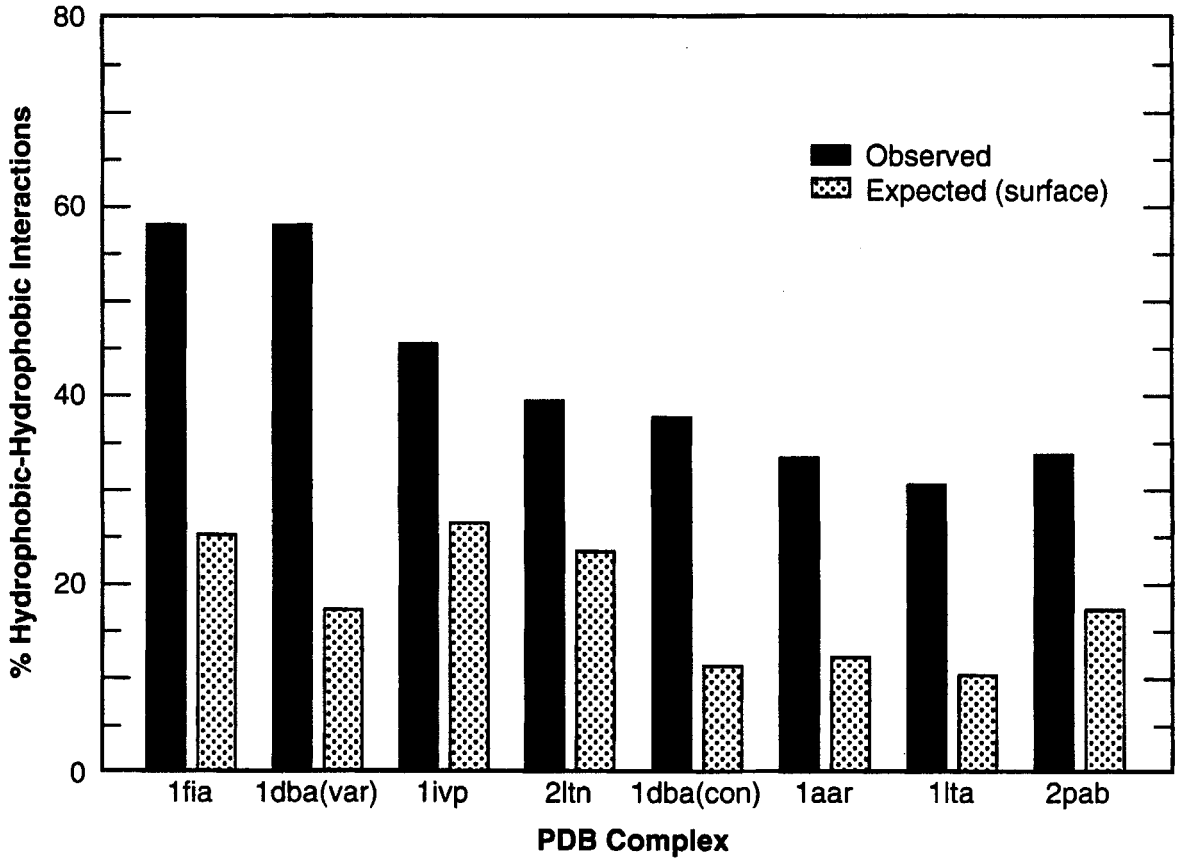


FIGURE 1. A histogram of the observed and expected hydrophobic interactions in the interfaces of the eight examples discussed here. The percentage of the observed is calculated as (*HH* + *HA* + *AH* + *AA*)/(*HH* + *HP* + *HA* + *PH* + *PP* + *PA* + *AH* + *AP* + *AA*), where the two-character symbol stands for the count of the different types of interactions between two chains. The criterion for an interaction between two amino acids across the interface is that the distance between one atom from each is less than, or equal to, the sum of their van der Waals radii + 0.5 Å. *H* stands for the hydrophobic amino acids; *P* is for hydrophilic, and *A* is the label used for amphiphilic ones. The amino acids belonging to each category are enumerated in the text.

its ASA is below this threshold, it is classified as a buried amino acid. This procedure has been carried out both for the single monomers and when in association in the complex. The percentages of hydrophobic amino acids that have been classified as surface residues in single chain structures and buried ones in the respective complex are plotted in Figure 2, along with the expected values. The latter are calculated from the composition of the surfaces of the monomers in each of the examples. A consistent trend of a larger than expected number of hydrophobic amino acids that have become buried after subunit association is observed. The four helix bundle of the Fis protein and the β-barrel between the variable domains of the immunoglobulin demonstrate the strongest hydrophobic effect using both parameters, the percent hydrophobic interactions and hydrophobic amino acids whose status have changed from surface to interior (Figures 1 and 2).

Loops are frequently composed of hydrophilic amino acids and are generally found on protein surfaces. If the underlying

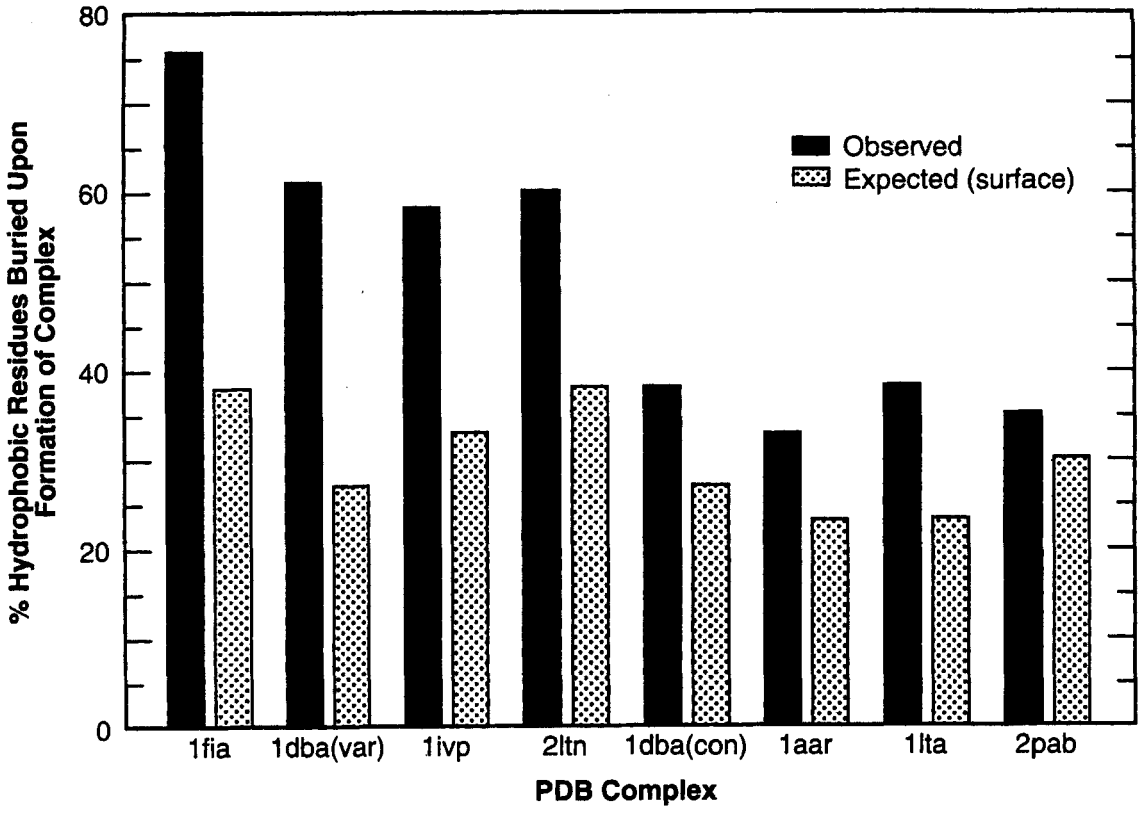


FIGURE 2. A histogram of the percentage of hydrophobic amino acids that were on the surface of the chains and subsequently buried in the interface during subunit association. This observed percentage of hydrophobic residues whose status changed from surface to interior is calculated as $(H_a + H_b)/(H_a + H_b + P_a + P_b + A_a + A_b)$, where H, P, and A stand for the count of status changed for hydrophobic, hydrophilic, and amphiphilic residues, respectively, in proteins a and b. The expected value is calculated using the composition of the amino acids on the surfaces of the two chains.

physical principles of protein-protein binding are similar to those of protein folding, then one would not expect to find loops in the midst of the interacting surfaces of the protein chains, but rather at their edges. This proposition appears to hold in the survey we have made so far.

V. HYDROGEN BONDS AND ION PAIRS AT THE INTERFACES

It has long been known that the main driving force for folding globular proteins is to pack the hydrophobic side-chains in the interior of the molecule (Dill, 1990; Rose and Wolfenden, 1993). The results presented above support the notion that hydrophobic side-chains are a critical driving force in the association of protein molecules as well. Nevertheless, in addition to the hydrophobicity-driven folding and molecular recognition and association processes, hydrogen bond formation between backbones, between backbone and side-chain, and between side-chains, as well as strong ionic interactions such as in ion pairs (salt bridges), might play an important role in stabilizing the folded architectures. Thus, we have ana-

lyzed the interfaces presented in Plates 1 and 2 for hydrogen bonds (using the definition of Kabsch and Sander, 1983) and for ion-pairs. Ion-pairs have been defined as attractive electrostatic interactions between oppositely charged groups. A distance of less than, or equal to, 4 Å has been used. The positively charged groups considered are nitrogen atoms in the side-chains of His, Arg, and Lys. The negatively charged groups taken into account are the oxygens in the side-chains of aspartic and glutamic acids. A total of nine ion-pairs have been detected in our eight examples presented above. This suggests that while ionic interactions may play a role in specific protein-protein interactions, their contribution to the overall stability of the subunit associations is not as large as that observed for the hydrophobic interactions. The contribution from hydrogen bonds was similarly relatively low. Rather surprisingly, in the barrel architecture of the variable and constant domains of the immunoglobulins (1 dba), no backbone-backbone hydrogen bonds between the light and heavy chains were detected.

VI. THE FOUR-HELIX BUNDLES

The four-helix bundles have been analyzed in detail in both the protein monomers (Presnell and Cohen, 1989; Harris et al., 1994; Lin et al., 1995) and in the protein-protein interfaces (Lin et al., 1995). The analysis has shown the bundles to be as frequent in the interfaces as in the monomers, despite the fact that interfaces are composed of at least two chains, suggesting that protein-protein associations and protein folding are driven by similar principles. Comparisons of the statistics in the two collections of bundles, from the monomers and from the interfaces, enable gaining some insight into the mechanisms that are involved in bundle formation (Lin et al., 1995).

First, in both the monomers and the interfaces a favorable pathway for the formation of the four-helix bundles is indicated: the helices nucleate first in parallel pairs before coupling into a bundle. For the case of the interfaces such a two-step process is logical, as the two-helix nucleation step would first take place in the subunit, and then followed by subunit associations. Second, although the dipole-dipole interactions between the helices can be a stabilizing factor, it cannot be a dominant force. This is indicated by the variations in the helical dipole alignment between the monomer and the interface bundles. Third, the inter-helical connections are likely to affect the coupling of the helices. This is indicated by inspection of the interface bundles, showing a larger extent of disorientation of the helices in the intefaces than in the monomers. While energetics favors less contorted inter-helical connections, and evolution pressure favors shorter connections, both imply a preference toward up-and-down helix pairs, where the connections exist. Fourth, the bundles in both the monomers and the interfaces demonstrate concerted twist of the helices, leading to compact packing. Fifth, inter-helical stacking is quite well conserved in the monomers. The stacking is manifested by the compactness and the regular alignment of the amino acids on the helical surfaces. This, however, is not the case for the four-helix bundles in the interfaces, where no such stacking regularity is observed. Nevertheless, the four-helix bundles are as popular in the interfaces as in the monomers. To account for their frequent occurrence, despite the lack of inter-helical stacking, Lin et al. (1995) have proposed a "hard" and a "soft" interface association models. In these models, the subunits are more rigid, and are preformed, while the monomers are folded from more

flexible fragments. In the first scenario, the two moieties of the bundle, contributed by the two subunits, achieve compensatory stabilizing interactions for the subunit association, obtaining a local minimum, which is geometrically far from the global minimum conformation. In the second scenario, while the helices are restricted they can still adjust, and their residues at the interface are flexible enough to find near optimum interactions. In this model the stacking interactions are the outcome of the conformational space that the helices can explore. Here the folding pathway may be detected via the packing patterns of the helices of the bundles (Lin et al., 1995).

VII. DERIVATION OF A DATASET OF PROTEIN-PROTEIN INTERFACES

The task of the derivation of a structurally non-redundant dataset of protein-protein interfaces is far more complex than that of a structurally distinct monomer dataset. Interfaces are inherently composed of single, isolated residues that are in contact with residues in the opposing chain, and residues that are in their vicinity, in the supporting matrix. Thus, not only are interfaces composed of two independent chains, but, in addition, only pieces of each are actually included in the interfaces. These fragments, and the isolated amino acids constituting the interfaces, are arranged in the interfaces in different directionalities in the different proteins. Constructing a dataset of interfaces thus requires an algorithm that can compare proteins in a manner that is entirely independent of the order of the amino acids in the chains. Such an algorithm would enable not only construction of an interface dataset, but also its comparisons with the structures of the monomers, to unravel architectural motifs that recur both in the interfaces and in protein cores. In addition, architectural motifs in the interfaces, and in families of interfaces, can be investigated.

The frequently adopted procedure for constructing a dataset of monomers is first to carry out sequence comparisons (Boberg et al., 1992; Pascarella and Argos, 1992; Holm et al., 1993; Orengo et al., 1993; Orengo, 1994; Fischer et al., 1995b). This is followed by comparisons of the structures. Most algorithms that are designed for the comparisons of structures, are either based on string editing techniques (e.g., Taylor and Orengo, 1989), or they cluster similar transformations obtained in matching the structures of fragments of consecutive amino acids (Alexandrov et al., 1992; Vriend and Sander, 1991). Another approach has been to match secondary structure elements, in a manner that is entirely independent of their order (e.g., Mitchell et al., 1990; Grindley et al., 1993; Mizuguchi and Go, 1995; Alesker et al., 1995). Algorithms that are an adaptation of string editing approaches are amino acid order dependent, that is, they cannot match amino acids in a manner that is independent of their order in the polypeptide chain. Thus, even though the coordinates of the amino acids in two proteins may superimpose better if their order in the chain is disregarded, string editing techniques would be unable to detect such a match. In the fragment matching based approach, while order-independent matches are allowed between the fragments, the order of the amino acids in the fragments is kept, and no insertions or deletions there are allowed. Structural matching algorithms based on either of these approaches are thus entirely unsuitable for the task of comparisons of interfaces, composed of isolated amino acids, and where the order of the amino acids in the two chains, and in the pieces that are included in the interfaces should not be considered. A few years ago we intro-

duced a computer vision-based approach that is ideally suited to handle such a task (Nussinov and Wolfson, 1991; Bachar et al., 1993; Fischer et al., 1994). We shall give a somewhat informal description of our technique for the structural comparison of C_α atoms between two interfaces. The exact details appear in our above-mentioned references.

Our approach views protein structures (in this case interfaces) as collections of points (atomic coordinates) in space trying to detect some subsets of these points that have similar geometric shape (constellations). One would like to attach to each point in the structures some geometric attribute that could be compared between the two structures to detect 'almost' similar points. A natural geometric attribute of a point in space are its three-dimensional coordinates. If both structures could be computed in some 'absolute' coordinate frame, it would be enough to compare the coordinates of the points in both structures and extract a maximal matching subset. Obviously, this is not the case. Both structures were computed in different coordinate frames and in order to superimpose them in a way that produces a maximal matching subset, some unknown rotation and translation is required. However, the idea of coordinate comparison should not be abandoned entirely. If one represents each structure redundantly in many 'natural' coordinate frames and finds an efficient way to simultaneously compare all the representations of one structure vs. all the representations of the other, then those pairs of representations that yield a 'large enough' fit provide candidate solutions for the structural comparison, namely, the rotation and translation that superimpose the coordinate frames of both representations result in a 'large enough' superposition of the individual atoms of both structures. The problem remains which coordinate frames to choose in each structure and how efficient such an algorithm can be. On the one hand, the coordinate frames (representations) should be chosen in a way that ensures that no 'large matches' are overlooked. On the other hand, an excessive number of such representations will adversely affect the efficiency of the algorithm. To address the first concern, one can choose any ordered triplet of C_α atoms in a protein (interface) structure and uniquely define a 3-D Cartesian coordinate frame on it (Nussinov and Wolfson, 1991). The coordinates of all the atoms (not belonging to the frame) are computed in it and stored in a look-up table (hash-table) as follows. The coordinates of each atom (in the given frame) form an address to the table, and the information stored there is the identity of the frame triplet (basis) and the current atom. This is performed for each triplet of atoms in the structure. If the structure has m atoms, the above *preprocessing* phase is of the order of m^4 computations. Now the other structure is processed and compared with the first one via the look-up table, namely, for each ordered triplet of atoms a coordinate frame is defined, and the coordinates of all the atoms are computed. Each such coordinate becomes an address to the look-up table, and all the (first structure) frames appearing in the appropriate entry score a 'vote'. Note that addressing an occupied entry means that the points have similar coordinates in the associated frames. If some frame in the first structure receives a relatively high score, it indicates that its superposition with its associated frame will result in superimpositioning of a large number of atoms that contribute to this score. Assuming that the second structure has n atoms, this *recognition* phase requires on the order of n^4 computations. Note, that the total number of computations is the sum of

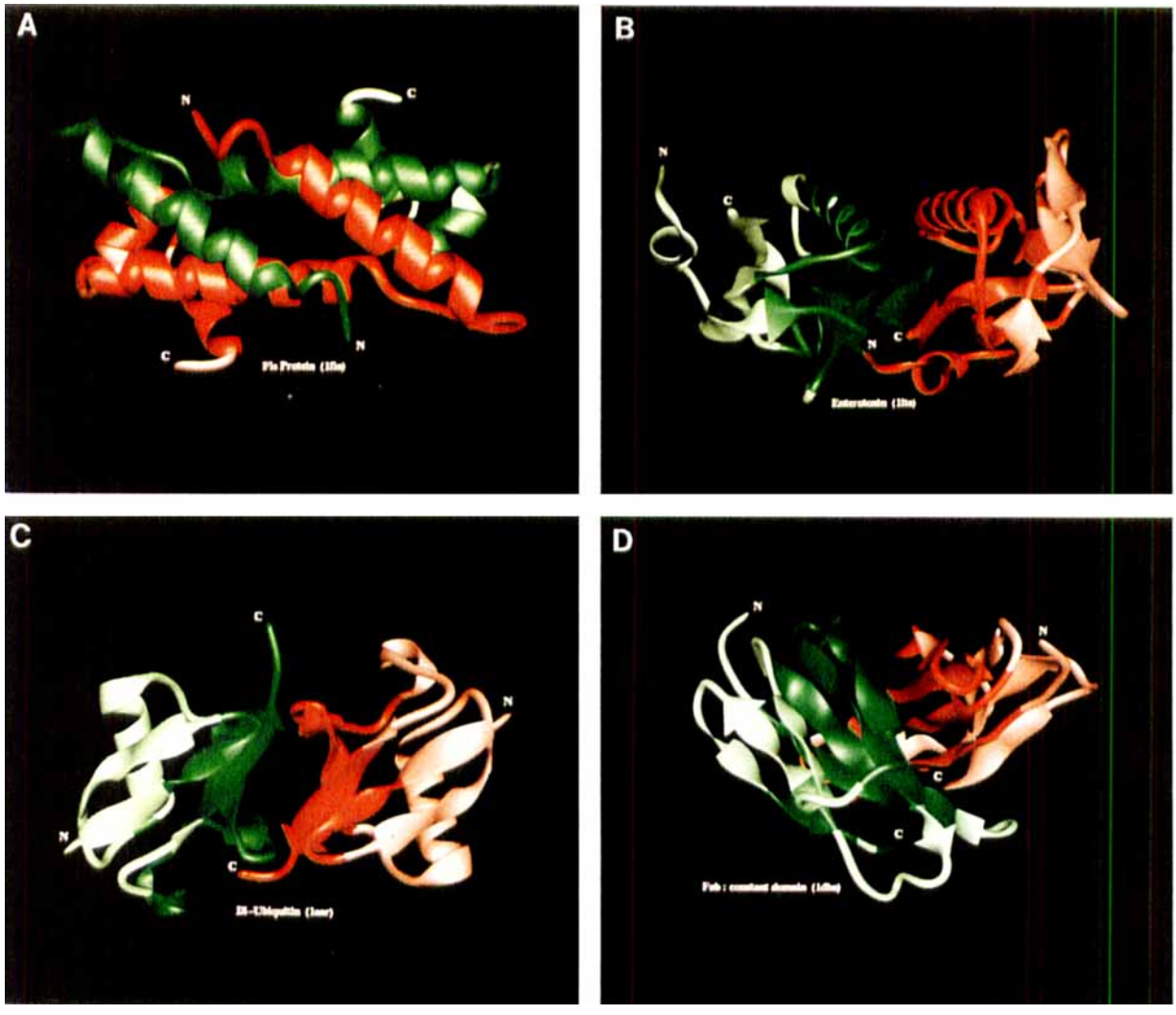


PLATE 1. Some examples of the types of motifs observed at the interfaces between two chains. The chains are easily distinguished by a distinctive color system, with protein backbone shown in ribbon representation (Midas plus, 1991). The darker colors (red and green) in each chain represent both the interfacing residues and the nearby ones. The lighter colors (pink and lighter green) represent the residues further away. The criteria used in their definition are outlined in the text. It is these darker-colored regions that are inspected for interface structural motifs. The interfacing residues are those having atoms whose distance across the interface is less than, or equal to, the sum of their van der Waals radii + 0.5 Å. The nearby residues are those whose C_a's are within a distance of at most 6 Å from a C_a of an interface residue.

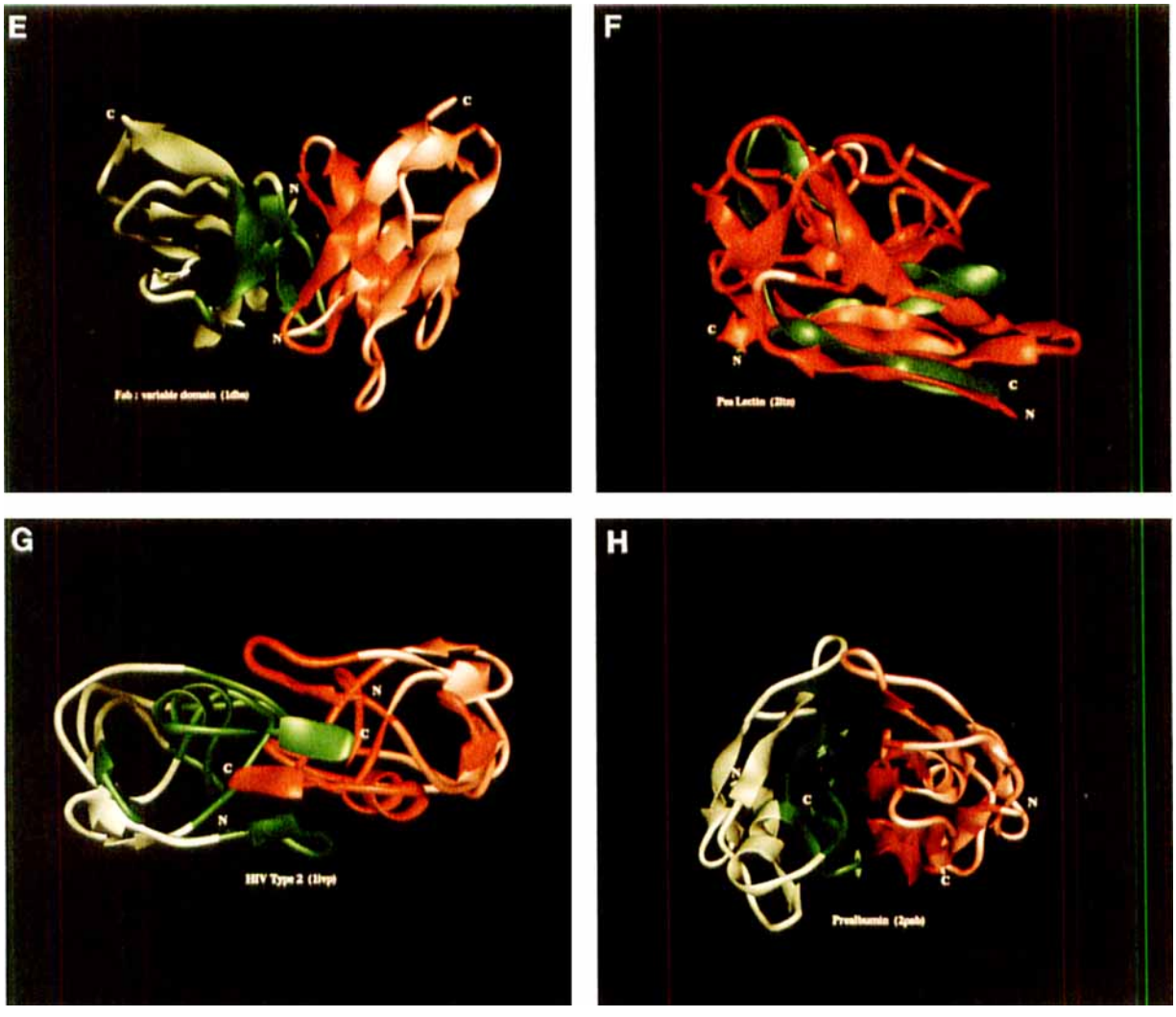


PLATE 1E, 1F, 1G, 1H.

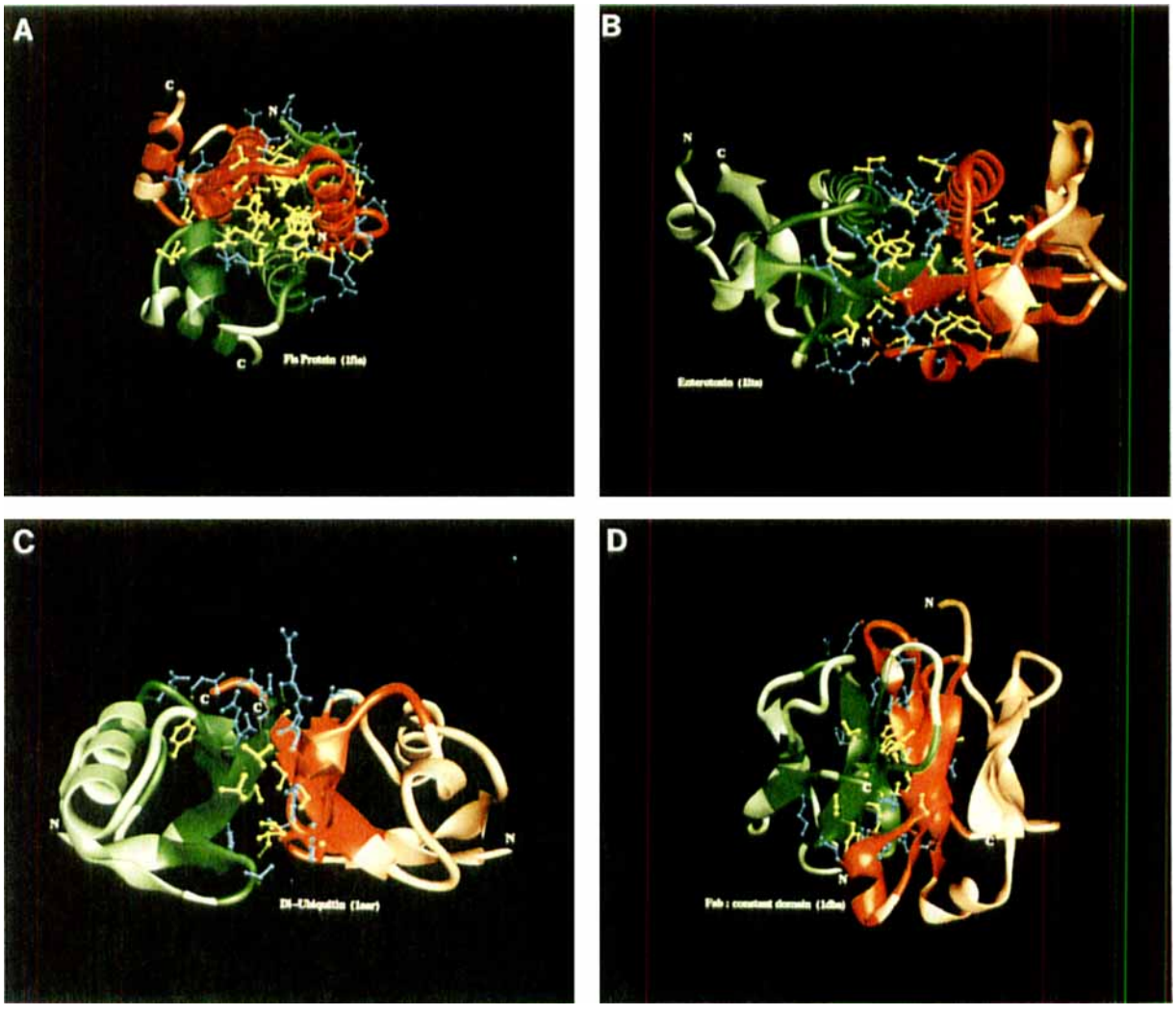


PLATE 2. Hydrophobic (yellow) and hydrophilic (cyan) side-chains at the two chain interface are drawn here, highlighting the hydrophobic interactions. The only side-chains drawn are those in which at least one atom from their corresponding amino acid interacts with an atom from the second chain. Interacting atoms are those whose distance is less than, or equal to, the sum of their van der Waals radii + 0.5 Å. The darker and lighter colors for each of the chains are as noted in the legend to Plate 1 and in the text.

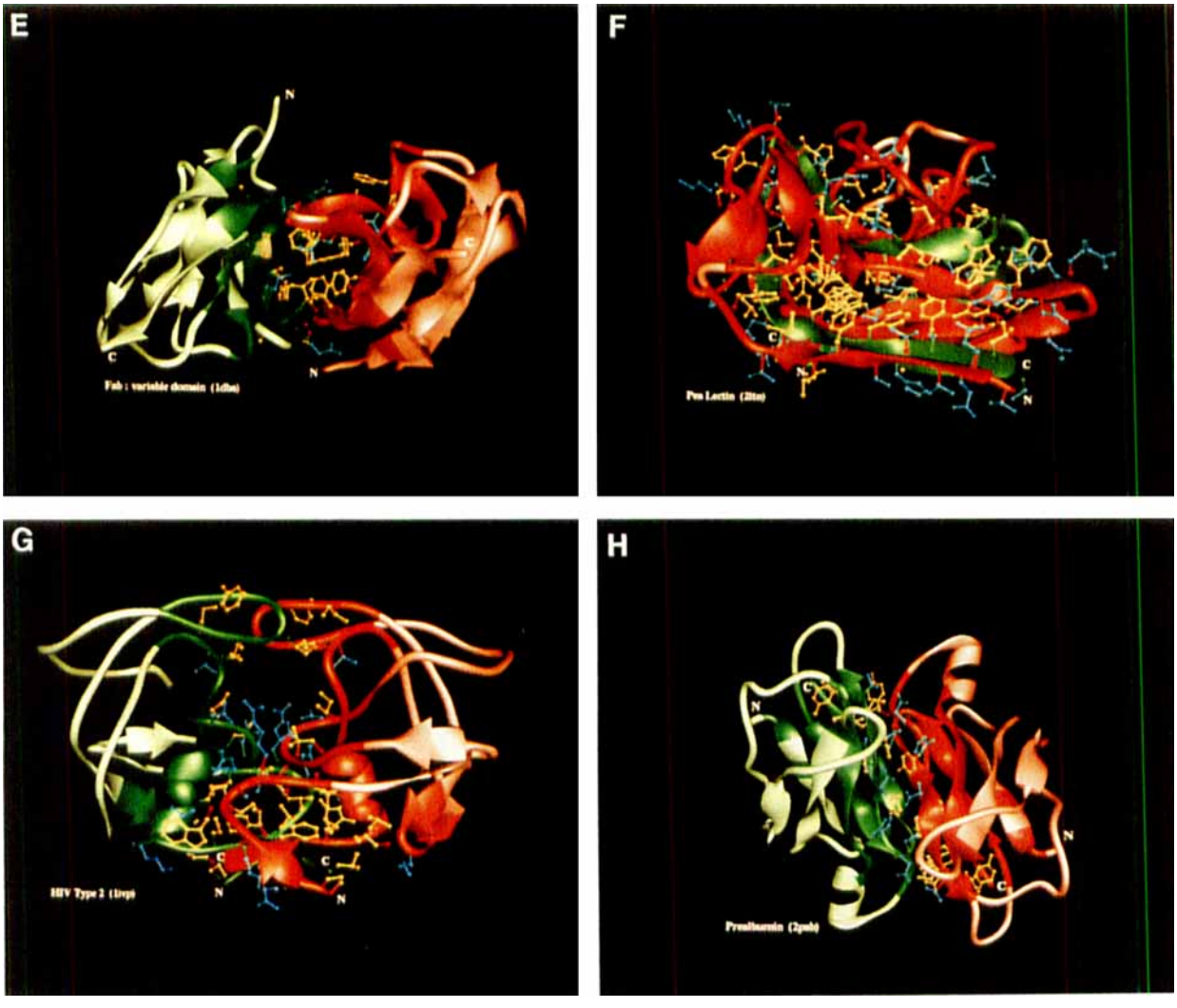


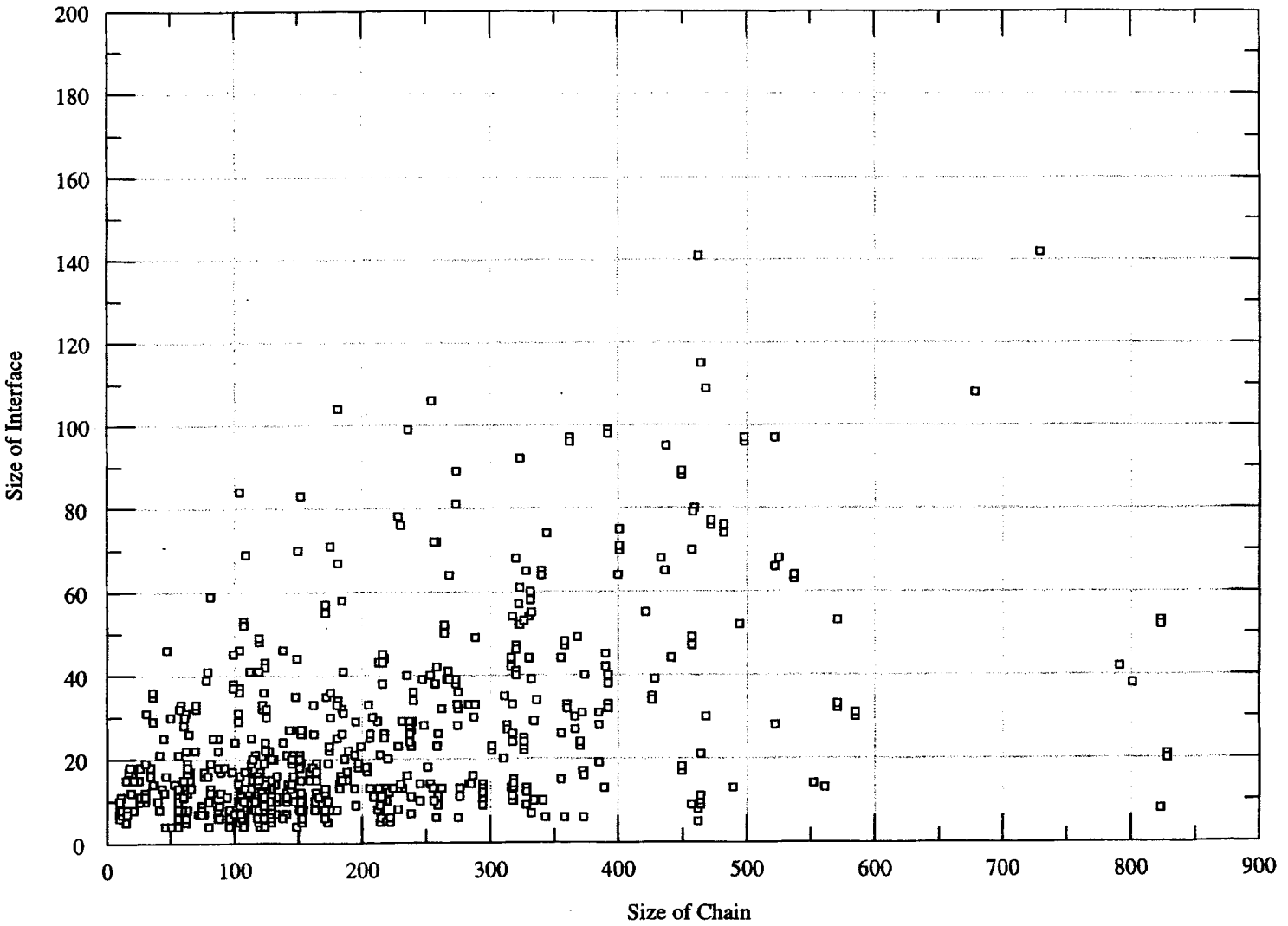
PLATE 2E, 2F, 2G, 2H.

the computations in both phases. Several enhancements can significantly speed-up this procedure (Bachar et al., 1993; Fischer et al., 1994). In particular, by a 'minute' exploitation of the sequence order, one might reduce the computational complexity of both phases to m^2 and n^2, respectively. This is achieved by defining coordinate frames only on triplets that are consecutive on the chain (Fischer et al., 1994). This leaves us with the order of m and n frames only. It is crucial to note that points not belonging to the current frame are totally unrestrained by connectivity constraints. One should also note that a unique reference frame can be defined on a single element of the peptide chain by using a triplet of atoms in such an element, for example, C_α, C_β, and N (Pennec and Ayache, 1994). In our implementation as described above we have used frames based on consecutive triplets, mainly to achieve better accuracy of computation.

After the 'voting' procedure is complete, one can superimpose the structures by computing the transformations between frames whose pairing received a high vote. This superimposition naturally aligns the atoms with 'almost' similar coordinates and might even suggest a fit between additional atoms. Recently, we have incorporated a biologically meaningful measure of similarity into this stage of the algorithm (Tsai et al., 1996c). For each matching pair of atoms in the superimposed structures, the matched/unmatched condition of the residues bordering a matched pair on both of its sides (in the polypeptide chain) is considered. Matches supported by neighboring matching pairs achieve a higher score. This allows an appreciable increase in the allowed RMS deviation of all the matching pairs, while, at the same time, not increasing the noise, resulting from randomly superimposed C_α atom-pairs, which happen to occupy similar positions in space. Interface atoms, which have no neighbors, achieve a full score. Thus, we exploit connectivity to improve the evaluation of intermediate results, where appropriate, without constraining our overall matching procedure.

Using this Geometric Hashing algorithm, and a fast, heuristic, clustering algorithm devised for this purpose (Tsai et al., 1996a), a dataset of representative protein-protein interfaces has been obtained. The dataset contains 351 entries. It has been generated following five iterative cycles of comparisons and clustering into families of homologous structures, with gradual relaxation of the parameters used for the clustering. At every cycle, the structure most resembling all other structures in its family has been chosen to represent the other family members. A detailed description of the generation of the dataset, and a listing of its members, has been given elsewhere (Tsai et al., 1996a). A full listing of the family members represented by each dataset entry as well as some of the characteristics of the dataset and its families has also been placed in the World Wide Web ("http://www-lmmb.ncifcrf.gov/~tsai").

A few of the attributes of the dataset of the protein-protein interfaces are shown here. As expected, in general, there is a direct correlation between the size of the protein monomer and the number of residues it contributes to the interface (Figure 3A); between the size of the protein monomer and the number of nearby noninteracting residues (i.e., those residues that are within the 6.0 Å radius from the residues interacting across the interface with the opposite monomer, see criteria for interface selection, above), which it contributes to the interface (Figure 3B); and between the number of interacting and noninteracting residues, which are included in the interface (Figure 3C). The scatter observed in Figure 3B, and



A

FIGURE 3. Some attributes of the interface dataset. (A) The number of interacting residues with respect to the size of each chain. (B) The number neighboring residues, fulfilling the 6 Å criterion with respect to the size of each chain. (C) The number of interacting residues with respect to the number of nearby's in the supporting matrix. The definitions of interacting residues, and of "nearby" ones are given in the text.

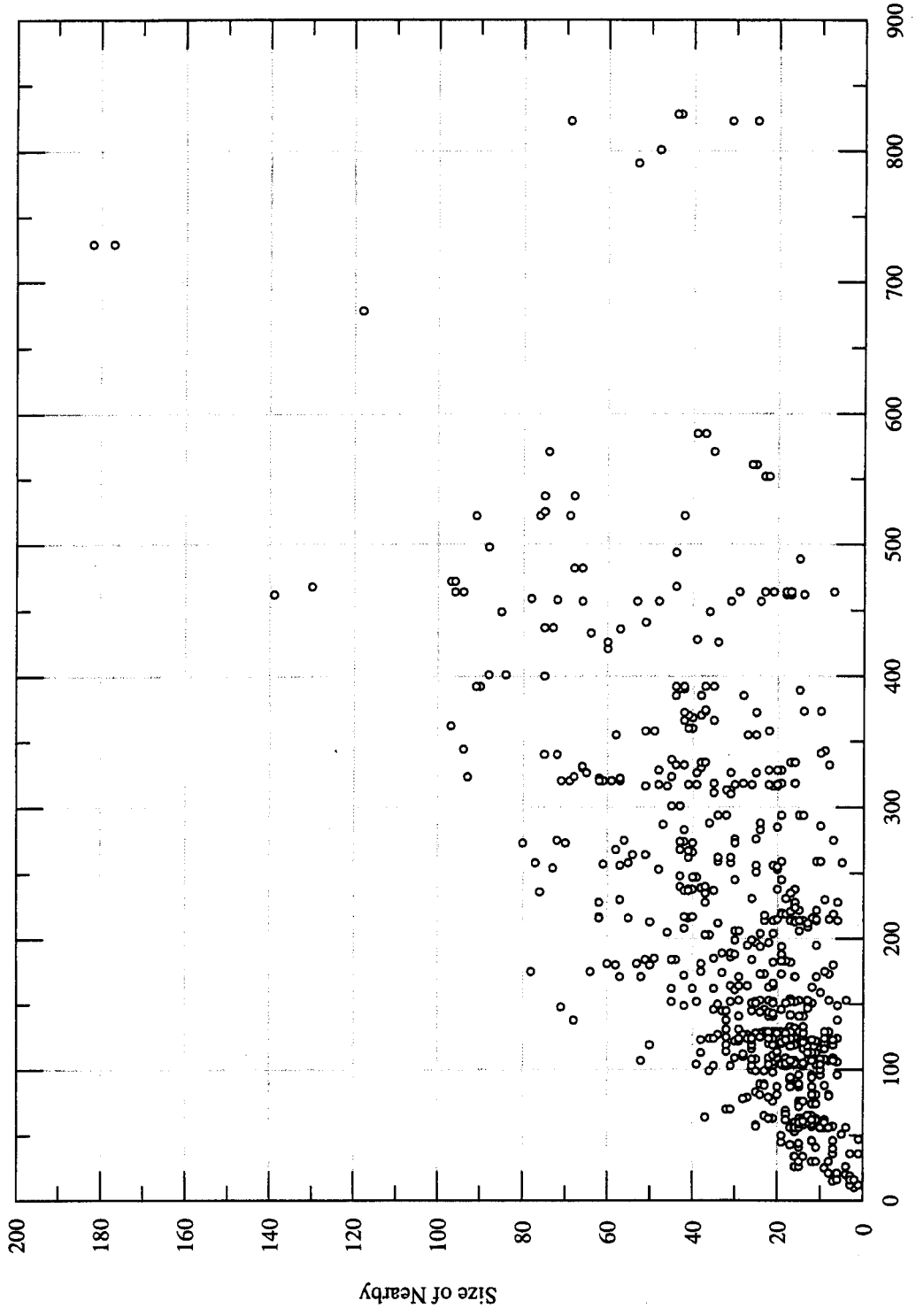


FIGURE 3B

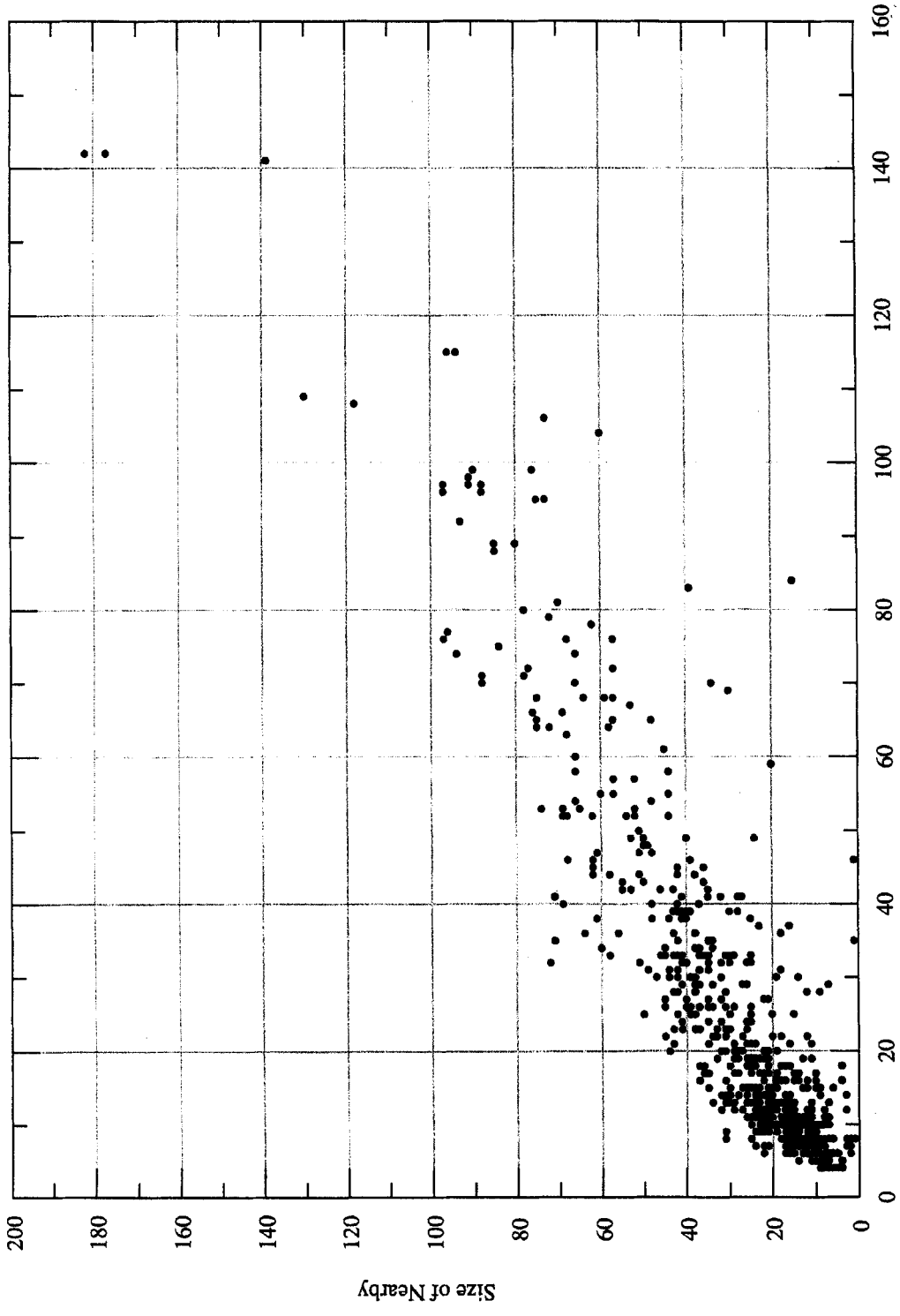


FIGURE 3C

particularly in Figure 3A, is due to the fact that while here we construct a two-chain interface for some proteins multi-chain interfaces are actually involved.

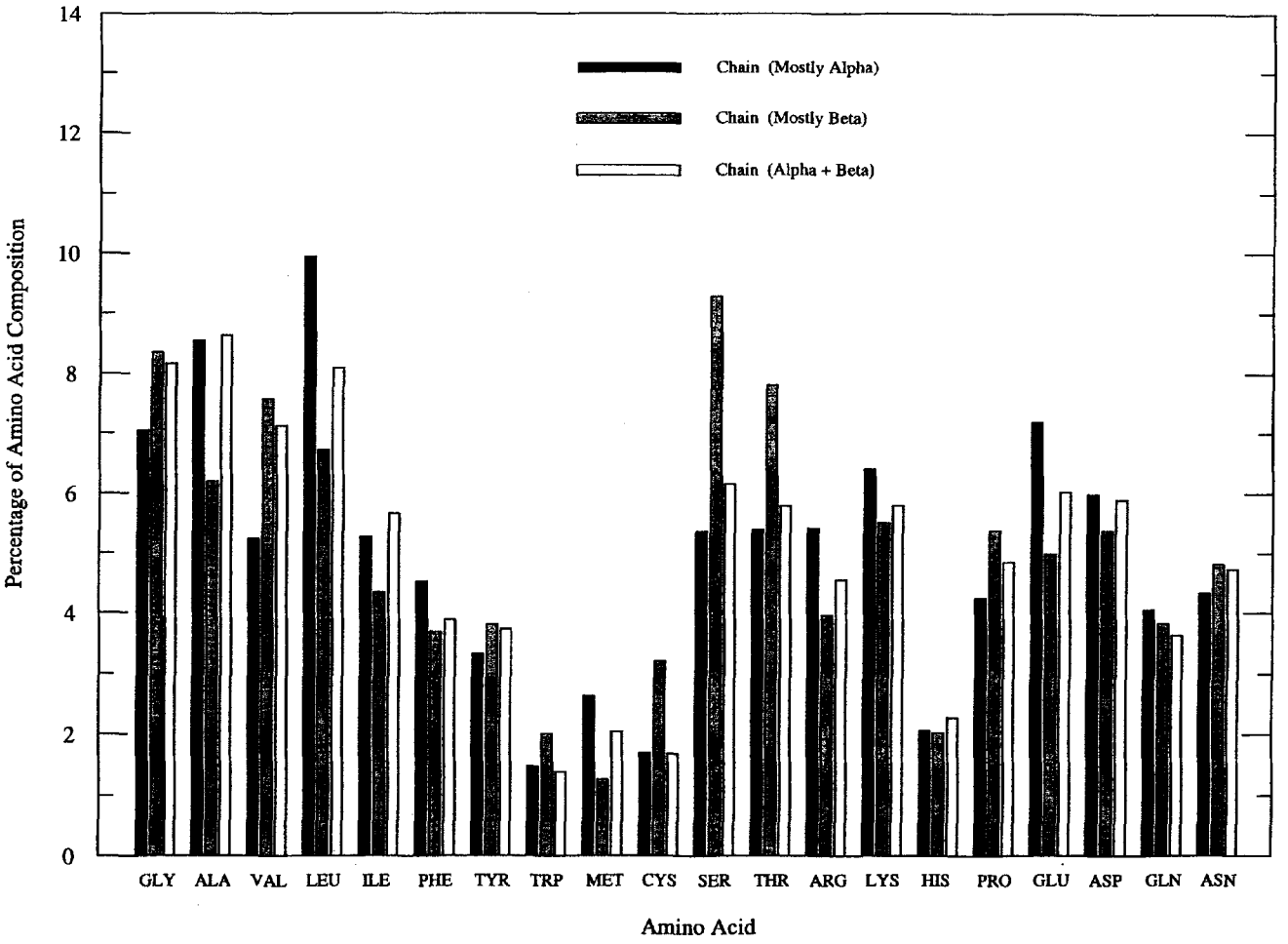
A similar procedure has also been adopted, and carried out, for the generation of a dataset of sequence- and structure-nonredundant dataset of protein monomers from all chains found in all entries of the PDB structural database. This dataset has been compared with the dataset of the interfaces in an all-against-all endeavor (Tsai et al., 1996b). Owing to the efficiency of our technique (a pair-wise comparison of two protein structures by the Geometric Hashing algorithm, takes on average 3 s on a Silicon Graphics workstation), such a comparison has been feasible. The dataset of the monomers includes 410 monomer structure entries.

It should be borne in mind that the dataset compiled and discussed here may not adequately represent all protein-protein interfaces. First, the PDB (Bernstein et al., 1977) is biased toward smaller, easier to crystallize proteins. Furthermore, some of the proteins, when in native form, do not form oligomers as they do after crystallization. While we have developed a procedure that to a certain extent may aid in removing interfaces that are suspected to be the outcome of the crystal packing force, and have applied it to the interface-dataset (see below), it is empirical, and is unlikely to remove all such crystal-packing interfaces.

Any comparison between structures involves some parameterization, and the procedure adopted here is no exception. Thus, on the one hand not all folds might be adequately represented; on the other hand, which is the more likely case, some redundancy might still exist. Furthermore, the thresholds that are employed, both in the Geometric Hashing and in the measures of similarity, affect the comparisons of the dataset of interfaces with the dataset of protein monomers, to detect recurring architectural motifs between the interfaces and the protein cores.

IX. ATTRIBUTES OF PROTEIN-PROTEIN INTERFACES: AMINO ACID AND SECONDARY STRUCTURE ELEMENT COMPOSITION

The recurrence of similar architectures in the interfaces and in single-chain proteins suggests that the compositions of the interfaces and of the chains are similar, both with respect to their amino acids and to the secondary structure elements. Figures 4A to c present the results of our analysis of the datasets of the single-chain monomers and of the protein-protein interfaces. Figure 4A presents the results for the chains; Figure 4B presents the results obtained for the interfaces, where the interfaces are composed both of the interacting residues and the supporting matrix. The latter include both the residues whose atom(s) interact across the interface with another chain, as well as residues whose C_α's are within 6.0 Å of a C_α of an interacting residue, as defined above. Figure 4C presents the results obtained in the analysis of only those residues that interact across the interface. In each of these histograms, the amino acid composition has been separately computed depending on whether the amino acid belongs to an α-helix, a β-strand, or a coil. Based on the fraction of the amino acids belonging to these secondary structure elements, the chain, or the interface, has been classified as belonging to the α-class; to the β-class, to the α+β-class, or to the coil class. No distinction has been made between the α/β class or the α+β, as the classification is based solely on composition. For this reason, here α+β implies both α+β and α/β.



A

FIGURE 4. Histograms of the distributions of the amino acids belonging to three classes of structures: those composed mostly of α-helices; those that consist mostly of β-strands and those that consist of α and β. Figure **A** displays the distribution of the composition of the amino acids in the single chains dataset; figure **B** depicts the distribution of the amino acids in the interfaces, which are composed of the interacting amino acids and their supporting matrix (i.e., the interacting residues, and their "nearby" ones, see text for definition). Figure **C** displays the distribution of the composition of the amino acids in the interacting amino acids. The distributions are quite similar in all three plots.

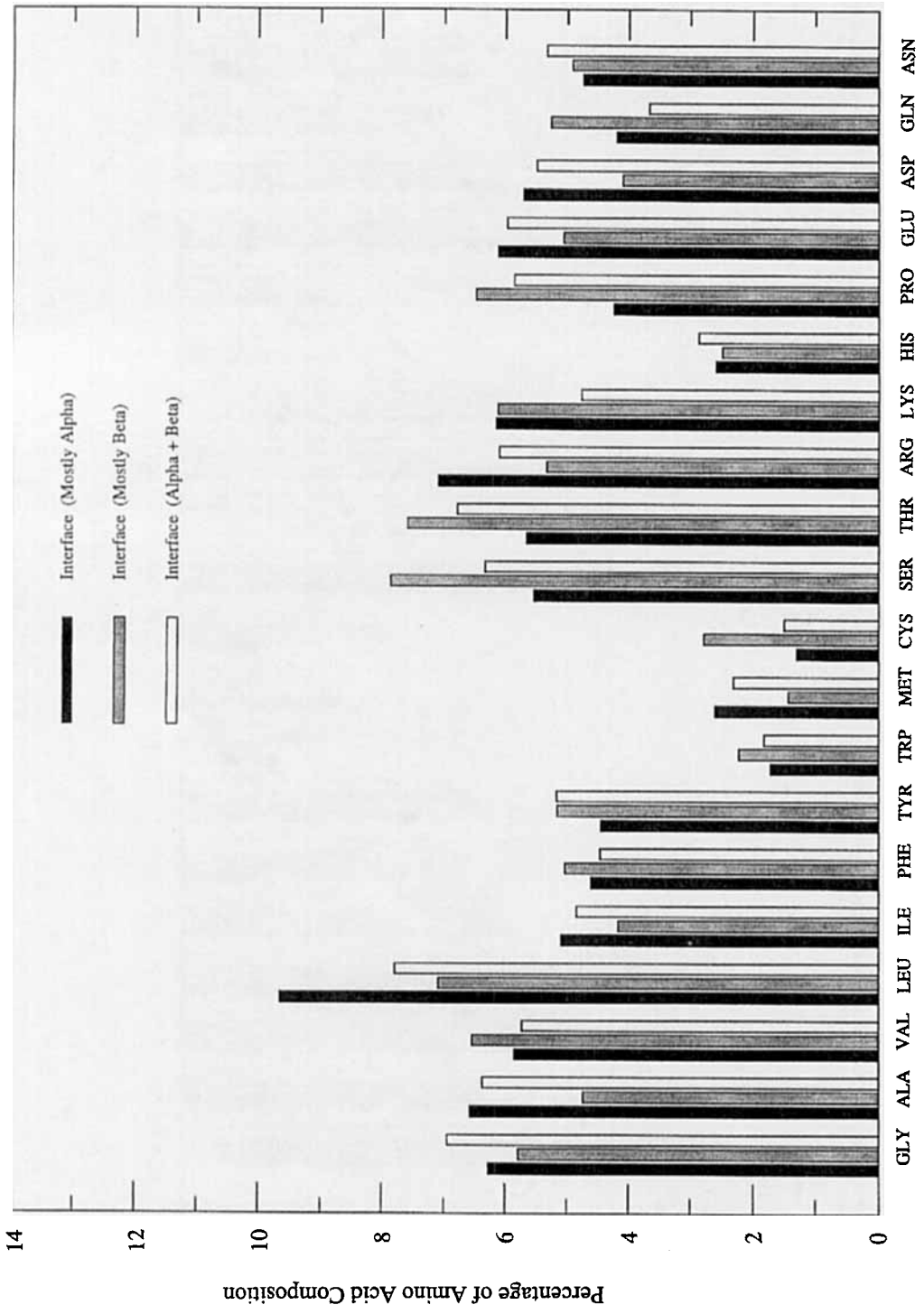


FIGURE 4B

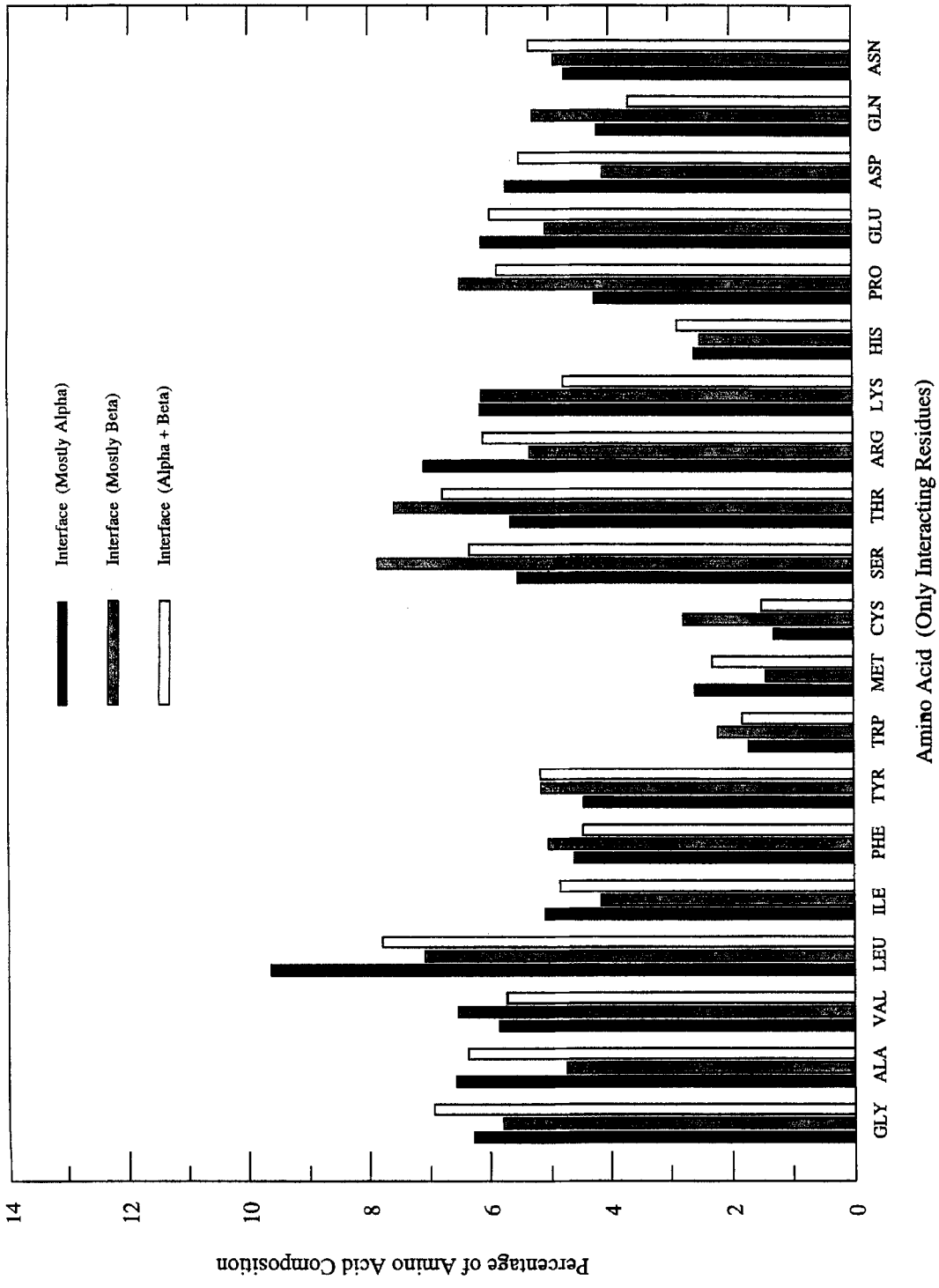


FIGURE 4C

To obtain these distributions, we have first assigned each amino acid, whether in the chain dataset or in the interface dataset to a secondary structure class. The assignment procedure is basically similar to that of Kabsch and Sander (1983), although it differs in some of its details. Briefly, we begin by generating the hydrogen atoms of the backbones. H-bonds are next located, and their energies are calculated, using the Kabsch and Sander procedure. Based on the energies, on the distances and on the O-H-N angles, the hydrogen bonds are classified as strong or as weak. If their energies are under −0.5 kcal/mol, the distance under 3.0 Å and the angle larger than 90°, the H-bond is considered strong; if the energy is between 0.1 and −0.5 kcal/mol, the distance under 3.5 Å and the angle larger than 90°, the H-bond is considered weak. The residues are next classified into donor, acceptor, or both categories, and a pattern of symbols is assigned. This pattern of symbols is scanned to assign the residues into α-helix, β-strand, or coil. The results of this assignment are generally very similar to those of DSSP (Kabsch and Sander, 1983).

To generate the plots of Figures 4A to C, the secondary structure assignments of the chains or of the interfaces are scanned, and the amino acids belonging to each type are counted. If the fraction of the amino acids found in coil (P_c) is larger than 0.7, and the fraction of the amino acids in α-helices (P_α) is less than 0.20 and the fraction of the amino acids in β-strands (P_β) is less than 0.20, the chain or interface is classified as belonging to the coil-class. If $P_\alpha/P_\beta > 4.0$, it is categorized as belonging to the α-class. Conversely, if $P_\beta/P_\alpha > 4.0$, it is categorized as belonging to the β-class. If it is in neither of the above, it is classified as belonging to the α+β class. Applying these criteria, we have obtained 135 interfaces in the α-class; 81 in the β-class; 144 in the α+β class; and 13 in the coil-class. Applying these to the single chain protein dataset, we have obtained 80 chains in the α-class, 64 in the β-class, 199 in the α+β class, and 15 in the coil-class. A comparison of these distributions of the secondary structure elements between the chains and the interfaces indicates that α-helices are more frequent in the interfaces than in chains, whereas the α+β class is more populated in chains than in interfaces. We hesitate, however, to draw any conclusions from these results, as they might simply reflect the similarity measure adopted for the construction of the datasets. As terminal residues (Tsai et al., 1996c) are more frequent in interfaces than in chains, the extent of similarity (or dissimilarity) between members of the representatives in the datasets might differ, affecting these statistics.

Inspection of Figures 4A–C reveals that the amino acid composition of each of the classes, in the chains (4A), in the interfaces which are composed of the interacting residues and the supporting matrix (4B), and of the interacting residues themselves (4C), are quite similar, although some minor variations exist.

X. PROTEIN-PROTEIN INTERFACES AND PROTEIN CORES: SIMILAR UNDERLYING PRINCIPLES AND DIFFERENCES

Three-dimensional architectures that are found in the cores of proteins recur at protein-protein interfaces also. Such secondary structure motifs are favorable not only in the interior of proteins, but play a critical role in the associations of their surfaces. The principles underlying the folding of the protein chains, the associations of domains, and the recognition and binding of different protein chains appear to be generally simi-

lar. This fact is hardly surprising. It has been known that proteins, and domains, often have similar folding units. While the structural details of these folding units vary, the general architectures are similar. These observations have already suggested that the reason for the recurrence of these motifs is some physical limitation on the set of folding patterns (Finkelstein and Ptitsyn, 1987). Because these motifs are favorable in the interior of the proteins, they are likely to be equally favorable at their interfaces. Stable stacking of their α-helical and their β-sheet layers is observed in the interfaces as well, with the polar backbone CO and NH groups forming hydrogen bonds within the secondary structure elements, and the hydrophobic side-chains from the secondary structure elements pointing toward each other. Polar loops are generally not observed inside the interacting protein surfaces. This arrangement fulfills the conditions required for thermodynamic stability of these compact motifs at the interfaces as well (Finkelstein and Ptitsyn, 1987).

These observations raise some questions and suggest several potential implications and applications. First, to what extent are the structures of the subunits similar when they exist as monomers and as multimers? It is easy to see that if the structures of the chains are somewhat different when separate, a conformational rearrangement may take place after binding to form favorable motifs. How extensive would the conformational rearrangement be is an open question. On the other hand, if the structures of the chains are relatively similar when they exist separately and in association, both the "part motif" and the "full motif" are conformationally stable. As the binding of the monomers involves segregation of hydrophobic groups from water, it is thermodynamically very favorable. The stability of the monomer and its conformational rearrangement may be related to the size of the subunit, to the hydrophobic interactions at its core, to the extent of the hydrophobic patch on its surface, and to the type of fold it forms at the interface. A particularly interesting example in this regard is the interlaced chains of the pea lectin, whose difference in size is appreciable. The larger of the two chains (181 amino acids) may form the motif shown in Plate 1F in the absence of the smaller chain and undergo conformational rearrangement during the process of interaction, allowing the insertion of the β-strands from the smaller chain. However, the smaller chain (47 amino acids) could hardly attain that structure by itself.

Second, the observations shown here are for protein-protein interfaces, where the chains are subunits associating into multimers. Associations of structural proteins may exhibit the same patterns. Whether such architectures are also likely to be found when the two proteins are enzyme-substrate associations is unclear. The nature of subunit-subunit interactions is different from that of the enzyme (catalyst) and its substrate, where the enzyme-substrate complexes should not be too stable so as to allow dissociation. In the latter cases, hydrophobic interactions may not be as strong. Analysis of the hydrophobic effect in subunit-subunit, receptor-ligand and enzyme inhibitor complexes, indicates that hydrophobicity plays a lesser role in receptor-ligand interactions than in subunit-subunit or enzyme-inhibitor ones (Tsai et al., 1996b). Our investigations of enzyme-ligand, as well as of antibody-antigen crystal-complexes indicate that they associate chiefly via loops. Loops often contain hydrophobic and hydrophilic elements. This ingenious design by nature allows the loops to exist both in the complex and in the separate states, avoiding major conformational rearrangements of the receptors that might otherwise be required. However, we cannot exclude the possibil-

ity that some types of motifs may occur at these interfaces.

Despite these overall similarities, the differences between protein folding and protein-protein recognition should not be overlooked. All too often these processes are considered practically identical, and considerations from the first are applied to the second. A frequently encountered example is applying inter-atom, or inter-residue, potential functions, derived from statistics of single chain protein structures, to binding and to considerations in ligand design. Solution measurements of binding constants of protein folding (e.g., Xie and Freire, 1994) differ from those of protein-protein binding (e.g., Horton and Lewis, 1992, and references therein). Indeed, the major difference in the fitting of the parameters between Horton and Lewis, who addressed protein associations, and Eisenberg and McLachlan (1986), who treated protein folding, is in the fitting of the hydrophilic pairs. Our analysis of the motifs indicates that while there is an overall similarity between chains and interfaces (Lin et a., 1995; Tsai et al., 1996a,b), the details differ (Lin et al., 1995). Analysis of the interactions in our interface dataset reveals that, whereas the hydrophobic effect has a dominant role in recognition, it is not to the same extent as in protein folding. On the other hand, hydrogen bonds and ion pairs may be more important in binding than in folding. It is the manifestation of such differences that are particularly revealing in investigations of folding and of recognition.

XI. INSPECIFIC INTERACTIONS MINIMIZE THE EFFECT OF SEQUENCE VARIABILITY

Inspection of the architectural motifs shown here demonstrates that they are largely composed of inspecific interactions. Their stabilization stems from hydrogen bonds between the backbones of β-strands, whether in parallel or anti-parallel orientation, and from the hydrophobic interactions between non-polar side-chains. This type of interaction, which is manifested in protein cores, minimizes the effect of sequence variability. In this regard, it is interesting to note that the crystal structure of a streptococcal protein G domain bound to a Fab fragment has shown that an outer β-strand in the protein G domain extends the β-sheet of the constant heavy chain of the immunoglobulin (Derrick and Wigley, 1992; 1994). As Derrick and Wigley note, this inspecific interaction, via hydrogen bonding of the backbones, between the anti-parallel β-strands at the interface, provides an ingenious solution to the problem of maintaining high affinity of protein G for different IgG molecules. The complex is further stabilized by interactions of the hydrophobic residues between the two chains. Backbone-backbone interactions across the interface, whether between β-strands or between loops, is thus an attractive solution to the problem of sequence variability. A similar mode of protein-protein interactions has been observed in the interface of the PapD chaperone with the pilus subunit from *E. coli* (Kuehn et al., 1993). Main-chain hydrogen bonds between parallel β-strands and contacts between the hydrophobic residues in the peptide, corresponding to the carboxy-terminus of the pilus subunits, and the chaperone, stabilize the complex. In addition, hydrogen bonds to two invariant residues, Arg and Lys, further contribute to the binding. Examination of the types of motifs at the interfaces with respect to the type of interactions that are involved addresses the question of how general is this solution for those proteins that bind to many different molecules. Side-chains involved in these interactions may be inspected with regard

to their conservation. As hydrophobic interactions are inspecific by their very nature, they provide both an added, substantially stabilizing means for these protein-protein interfaces, as well as an alternative to variable sequences demonstrating multiple, inspecific, binding. Such interactions might be particularly dominant in the absence of backbone hydrogen bonding, such as between α-helices. These motifs and interactions may be compared with others found in our dataset of protein interfaces. The availability of very fast, secondary structure element sequence-order independent comparison techniques for detection of motifs composed of α-helices, β-strands, or loops (e.g., Mitchell et al., 1990; Grindley et al., 1993; Alesker et al., 1996) facilitate such an undertaking. In particular, such inspecific binding may be the way poly-reactive antibodies bind to their ensemble of antigens. Protein G — Fab and PapD chaperone — pilus associations provide some examples.

XII. CONCLUSIONS AND IMPLICATIONS TO PROTEIN-PROTEIN (INHIBITOR) DOCKING

Although the architectures of the motifs at the interfaces and in protein cores are similar, the details of the structures vary. As the actual amino acid sequences differ, it is probably their general, crude features that specify the three-dimensional fold. The general similarity between protein cores and interfaces is consistent with some additional elegant studies of the binding surfaces. In order to estimate the interface complementarity in the antigen-antibody docking, Walls and Sternberg (1992) computed the average packing density over all atoms and the interface packing density, which is the average for the surface atoms. Their results indicate that the differences between the global and interface packing densities are much smaller than the standard deviation for all structures. This demonstrates that the interface region is as tightly packed as the protein core, suggesting that the interface region exhibits the steric complementarity observed in protein cores (Chothia and Janin, 1975). More recently, Young et al. (1994) have shown that by applying the hydrophobicity scale derived from single chain protein structures (Miyazawa and Jernigan, 1985) to protein surfaces, they have been able to successfully locate binding sites on protein surfaces. This reinforces the notion that the hydrophobic interactions in protein cores resemble those at their interfaces. One may thus postulate that while the inspecific hydrophobic interactions serve to stabilize the bound protein molecules, the orientation of one chain with respect to the other is determined by the internal architecture, which maximizes the van der Waals favorable complementarity. Computational approaches to the protein-protein docking problem may benefit from such considerations. One may scan protein surfaces searching for potential complementary "part-motifs". These might be favorable binding orientations. Such considerations may also be useful in the design of protein ligands.

The formation of dimeric (or oligomeric) proteins can be considered as a two-state or a three-state reaction (Neet and Timm, 1994). In the former, the conformations of the unassociated folded polypeptide chains are unstable, with the monomer not found in its native conformation in significant amounts in solution. In such a model, the folded monomeric chain is stabilized by its quaternary interactions. Alternatively, in the three-state model, the polypeptide chain folds into a stable monomer conformation. The monomers subsequently assemble into an oligomeric protein.

Architectural motifs have already been shown at protein-protein binding interfaces. Indeed, some are textbook examples (Branden and Tooze, 1991). Exhaustive comparisons carried out between all the nonredundant structures of the single chain proteins and all chain-chain interfaces compiled in the structural database (Bernstein et al., 1977) by our highly efficient, three-dimensional amino acid sequence-order independent comparison approaches (Nussinov and Wolfson, 1991; Fischer et al., 1994) show that this is a general phenomenon. Interfaces appear to possess the hallmarks of globular proteins, namely, a core of non-polar residues, compactness and the internal, motif-fold architecture. Nevertheless, the differences in the motifs and in the interactions should not be overlooked.

We thank Drs. Robert Jernigan and Jacob Maizel and Dong Xu for helpful discussions, encouragement, and interest. We thank the personnel at the Frederick Cancer Research and Development Center for their assistance. The research of R. Nussinov has been sponsored by the National Cancer Institute, DHHS, under Contract No. 1-CO-74102 with SAIC. The research of H. J. Wolfson has been supported in part by a grant from the *Israel Science Foundation* administered by the *Israel Academy of Sciences.* The research of R. Nussinov in Israel has been supported in part by grant No. 91-00219 from the BSF, and by a grant from the *Israel Science Foundation* administered by the *Israel Academy of Sciences.* The contents of this publication do not necessarily reflect the views or policies of the DHHS nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

Alesker, V., Nussinov, R., and Wolfson, H. 1996. Detection of non-topological motifs in protein structures, *Prot. Eng.* in press.

Alexandrov, N. N. and Go, N. 1994. Biological meaning, statistical significance, and classification of local spatial similarities in non-homologous proteins. *Prot. Sci.* **3:**866–875.

Alexandrov, N. N., Takahashi, K., and Go, N. 1992. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225:**5–9.

Arevalo, J. H., Stura, E. A., Taussig, M. J., and Wilson, I. A. 1993. Three-dimensional structure of an anti-steroid Fab' and progesterone-Fab' complex. *J. Mol. Biol.* **231:**103–118.

Argos, P. 1988. An investigation of protein subunit and domain interfaces. *Prot. Eng.* **2:**101–113.

Bachar, O., Fischer, D., Nussinov, R., and Wolfson. H. 1993. A computer vision-based technique for 3-D sequence independent structural comparison of proteins. *Prot. Eng.* **6:**279–288.

Banner, D. W., Kokkinidis, M., and Tsernoglou, D. 1987. Structure of the col rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196:**657–675.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112:**535–542.

Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rerat, B., and Rerat, C. 1978. Structure of prealbumin, secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Å. *J. Mol. Biol.* **121:**339–356.

Boberg, J., Salakoski, T., and Vihinen, M. 1992. Selection of representative set of structures from Brookhaven Data Bank. *Proteins* **14:**265–276.

Branden, C. and Tooze, J. 1991. *Introduction to Protein Structure,* Garland Publishing, Inc., New York and London.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. 1983. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.* **4:**187–217.

Cherfils, J. and Janin, J. 1993. Protein docking algorithms: simulating molecular recognition. *Curr. Opin. Struct. Biol.* **3:**265–269.

Chothia, C. 1992. One thousand families for the molecular biologist. *Nature (London)* **357:**543–544.

Chothia, C. and Janin, J. 1975. Principles of protein-protein recognition. *Nature (London)* **256:**705–708.

Clore, G. M., Omichinski, J. G., Sakaguchi, K., Zambrano, N., Sakamoto, H., Appella, E., and Gronenborn, A. M. 1994. High-resolution structure of the oligomerization domain of p53 by multidimensional NMR. *Science* **265:**386–391.

Connolly, M. 1986. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interfaces. *Biopolymers* **25:**1229–1247.

Cook, W. J., Jeffrey, L. C., Carson, M., Chen, Z., and Pickart, C. M. 1992. Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). *J. Biol. Chem.* **267:**16467–16471.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in protein. *J. Mol. Biol.* **195:**659–685.

Crippen, G. M. and Maiorov, V. N. 1992. How many protein folding motifs are there? *J. Mol. Biol.* **252:**144–151.

Derrick, J. P. and Wigley, D. B. 1992. Crystal structure of a streptococcal protein G domain bound to an Fab fragment. *Nature (London)* **359:**752–754.

Derrick, J. P. and Wigley, D. 1994. The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243:**906–918.

Dill, K. A. 1990. Dominant forces in protein folding. *Biochemistry* **31:**7134–7155.

Duncan, B. S. and Olson, A. J. 1993. Approximation and characterization of molecular surfaces. *Biopolymers* **33:**219–229.

Duncan, B. S. and Olson, A. J. 1993. Shape analysis of molecular surfaces. *Biopolymers* **33:**231–238.

Eisenberg, D. and McLachlan, A. D. 1986. Solvation energy in protein folding and binding. *Nature (London)* **319:**199–203.

Finkelstein, A. V. and Ptitsyn, O. B. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* **50:**171–190.

Fischer, D., Lin, S. L., Wolfson, H. J., and Nussinov, R. 1994. 3-D, sequence-order independent structural comparison of trypsin against the crystallographic database reveals active site similarities to subtilisin-like and sulfhydryl proteases: potential implications. *Prot. Sci.* **3:**769–778.

Fischer, D., Lin, S. L., Wolfson, H. J., and Nussinov, R. 1995a. A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **248:**459–477.

Fischer, D., Tsai, C.-J., Nussinov, R., and Wolfson, H. J. 1995b. A 3-D, sequence-independent representation of the protein data bank. *Prot. Eng.* **8(10):**981–997.

Goodsell, D. S. and Olson, A. J. 1990. Automated docking of substrates to proteins by simulated annealing, *Proteins: Struct. Funct. Genet.* **8:**195–202.

Grindley, H. M., Artymiuk, P. J., Rice, W., and Willet, P. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* **229:**707–721.

Harris, N. L., Presnell, S. R., and Cohen, F. E. 1994. Four-helix bundle diversity in globular proteins. *J. Mol. Biol.* **236:**1356–1368.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. 1993. A database of protein structure families with common folding motifs. *Prot. Sci.* **1:**1691–1698.

Horton, N. and Lewis, M. 1992. Calculation of the free energy of association for protein complexes. *Prot. Sci.* **1:**169–181.

Jackson, R. M. and Sternberg, M. J. E. 1995. A continuum model for protein-protein interactions: application to the docking problem. *J. Mol. Biol.* **250:**258–275.

Janin, J. and Chothia, C. 1990. The structure of protein-protein recognition sites. *J. Biol. Chem.* **265:**16027–16030.

Janin, J., Miller, S., and Chothia, C. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204:**155–164.

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

Kostrewa, D., Granzin, J., Stock, D., Choe, H.-W., Labahn, J., and Saenger, W. 1992. Crystal structure of the factor for inversion stimulation FIS at 2.0 Å resolution. *J. Mol. Biol.* **226:**209–226.

Kuehn, M. J., Ogg, D. J., Kihlberg, J., Slonim, L. N., Flemmer, K., Bergfors, T., and Hutgren, S. J. 1993. Structural basis of pilus subunit recognition by the PapD chaperon. *Science* **262:**1234–1241.

Lee, B. K. and Richards, F. M. 1971. The interpretation of protein structures. Estimation of static accessibility. *J. Mol. Biol.* **55:**379–400.

Lin, S. L., Tsai, C.-J., and Nussinov, R. 1995. A study of four-helix bundles: investigating protein folding via similar architectural motifs in protein cores and in subunit interfaces. *J. Mol. Biol.* **248:**151–161.

Lodi, P. J., Garrett, D. S., Kuszewski, J., Tsang, M. L.-S., Weatherbee, J. A., Leonard, W. J., Gronenborn, A. M., and Clore, G. M. 1994. High-resolution solution structure of the β chemokine hMIP-1 β by multidimensional NMR. *Science* **263:**1762–1767.

Milburn, M. V., Hassell, A. M., Lambert, M. H., Jordan, S. R., Proudfoot, A. E. I., Graber, P., and Wells, T. N. C. 1993. A novel dimer configuration revealed by the crystal structure at 2.4 Å resolution of human interleukin-5. *Nature* **363:**172–176.

Miller, S. 1989. The structure of interfaces between subunits of dimeric and tetrameric proteins. *Prot. Eng.* **3:**77–83.

Miller, S., Janin, J., Lesk, A. M., and Chothia, C. 1987. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196:**641–656.

Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willet, P. 1990. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212:**151–166.

Miyazawa, S. and Jernigan, R. L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18:**534–552.

Mizuguchi, K. and Go, N. 1995. Comparison of spatial arrangements of secondary structural elements in proteins. *Prot. Eng.* **8:**353–362.

Mulichak, A. M., Hui, J. O., Tomasselli, A. G., Heinrikson, R. L., Curry, K. A., Tomich, C.-S., Thaisrivongs, S., Sawyer, T. K., and Watenpaugh, . Human immunodeficiency virus type 2 protease mutant with Lys 57 replaced by Leu (K57L) complex with U75875 (NOA-HIS-CHA-PSI[CH(OH)CH(OH)]VAL-ILE-APY). PDB, 1993.

Neet, K. E. and Timm, D. E. 1994. Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. *Prot. Sci.* **3:**2167–2174.

Newcomer, M. E., Jones, T. A., Aqvist, J., Sundelin, J., Eriksson, U., Rask, L., and Peterson, P. A. 1984. The three-dimensional structure of retinol-binding protein. *EMBO J.* **3:**1451–1454.

Norel, R., Lin, S. L., Wolfson, H., and Nussinov, R. 1994. Shape complementarity at protein-protein interfaces. *Biopolymers* **34:**933–940.

Norel, R., Lin, S. L., Wolfson, H., and Nussinov, R. 1995. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. *J. Mol. Biol.* **252:**263–273.

Novotny, J., Bruccoleri, R., Newell, J., Murphy, D., Haber, E., and Karplus, M. 1983. Molecular anatomy of the antibody binding site. *J. Biol. Chem.* **258:**14433–14437.

Nussinov, R. and Wolfson, H. J. 1991. Efficient detection of motifs in biological macromol-

ecules by computer vision techniques. *Proc. Natl. Acad. Sci. U.S.A.* **88:**10495–10499.

Orengo, C. 1994. Classification of protein folds. *Curr. Opin. Struct. Biol.* **4:**429–440.

Orengo, C. A., Flores, T. P., Taylor, W. R., and Thornton, J. M. 1993. Identifying and classifying protein fold families. *Prot. Eng.* **6:**485–500.

Pascarella, S. and Argos, P. 1992. A data bank merging related protein structures and sequences. *Prot. Eng.* **5:**121–137.

Pennec, X. and Ayache, N. 1994. An $O(n^2)$ Algorithm for 3D Substructure Matching of Proteins. Proceedings of the First Int. Workshop on Shape and Pattern Matching in Computational Biology, 25–40.

Prasthofer, T., Phillips, S. R., Suddath, F. L., and Engler, J. A. 1989. Design, expression and crystallization of recombinant lectin from the garden pea (*Pisum sativum*). *J. Biol. Chem.* **264:**6793–6798.

Presnell, S. R. and Cohen, F. E. 1989. The topological distribution of four-alpha-helix bundles. *Proc. Natl. Acad. Sci. U.S.A.* **86:**6592–6596.

Rose, G. D. and Wolfenden, R. 1993. Hydrogen bonding, hydrophobicity, packing and protein folding. *Ann. Rev. Biophys. Biomol. Struct.* **22:**381–415.

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

Shrake, A. and Rupley, J. A. 1973. Environment and exposure to solvent of protein atoms. *J. Mol. Biol.* **79:**351–371.

Sixma, T. K., Kalk, K. H., van Zanten, B. A. M., Dauter, Z., Kingma, J., Witholt, B., and Hol, W. G. J. 1993. Refined structure of *Escherichia coli* heat-labile enterotoxin, a close relative of cholera toxin. *J. Mol. Biol.* **230:**890–918.

Taylor, W. R. and Orengo, C. A. 1989. Protein structure alignment. *J. Mol. Biol.* **208:**1–22.

Tsai, C.-J., Lin, S. L., Wolfson, H., and Nussinov, R. 1996a. Studies of protein-protein interfaces. I. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique, submitted.

Tsai, C.-J., Lin, S. L., Wolfson, H., and Nussinov, R. 1996b. Studies of protein-protein interfaces. II. Statistical analysis of the hydrophobic effect, submitted.

Tsai, C.-J., Lin, S. L., Wolfson, H., and Nussinov, R. 1996c. Techniques for searching for structural similarities between protein cores, protein surfaces and between protein-protein interfaces. *Techniques in Protein Chemistry* VII, in press.

Vriend, G. and Sander, C. 1991. Detection of common three-dimensional substructures in proteins. *Proteins: Struct. Funct. Genet.* **11:**52–58.

Wallqvist, A., Jernigan, R. L., and Covell, D. G. 1995. A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Prot. Sci.* **4:**1881–1903.

Walls, P. H. and Sternberg, M. J. E. 1992. New algorithm to model protein-protein recognition based on surface complementarity. *J. Mol. Biol.* **228:**277–297.

Xie, D. and Freire, E. 1994. Structure based prediction of protein folding intermediates. *J. Mol. Biol.* **242:**62–80.

Young, L., Jernigan, R. L., and Covell, D. G. 1994. A role for surface hydrophobicity in protein-protein recognition. *Prot. Sci.* **3:**717–729.