Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13–20, January 1996

Review

Principles of protein–protein interactions

Susan Jones and Janet M. Thornton

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England

ABSTRACT **This review examines protein complexes in the Brookhaven Protein Databank to gain a better understanding of the principles governing the interactions involved in protein–protein recognition. The factors that influence the formation of protein–protein complexes are explored in four different types of protein–protein complexes—homodimeric proteins, heterodimeric proteins, enzyme–inhibitor complexes, and antibody–protein complexes. The comparison between the complexes highlights differences that reflect their biological roles.**

1. Introduction

Many biological functions involve the formation of protein–protein complexes. In this review, only complexes composed of two components are considered. Within these complexes, two different types can be distinguished, homocomplexes and heterocomplexes. Homocomplexes are usually permanent and optimized (e.g., the homodimer cytochrome *c'* (1)) (Fig. 1*a*). Heterocomplexes can also have such properties, or they can be nonobligatory, being made and broken according to the environment or external factors and involve proteins that must also exist independently [e.g., the enzyme–inhibitor complex trypsin with the inhibitor from bitter gourd (2) (Fig. 1*b*) and the antibody–protein complex HYHEL-5 with lysozyme (3) (Fig. 1*c*)]. It is important to distinguish between the different types of complexes when analyzing the intermolecular interfaces that occur within them.

The division of proteins in the July 1993 Brookhaven Protein Databank (PDB) (4) into multimeric states is illustrated in Fig. 2. This distribution is biased as it reflects only those proteins whose structures have been solved and, therefore, probably overrepresents the small monomers. However, it is clear that trimers are relatively rare compared with tetramers and that the numbers of structures in the higher multimeric states fall markedly, with the obvious exception of the viral coat proteins, which contain high numbers (e.g., 60, 180, and 240) of subunits (see ref. 5).

Previous work has centered on two aspects of protein–protein recognition: the development of algorithms to dock two proteins together (6–8), and the structural characterization of protein–protein interfaces. Janin *et al.* (9), Miller (10), Argos

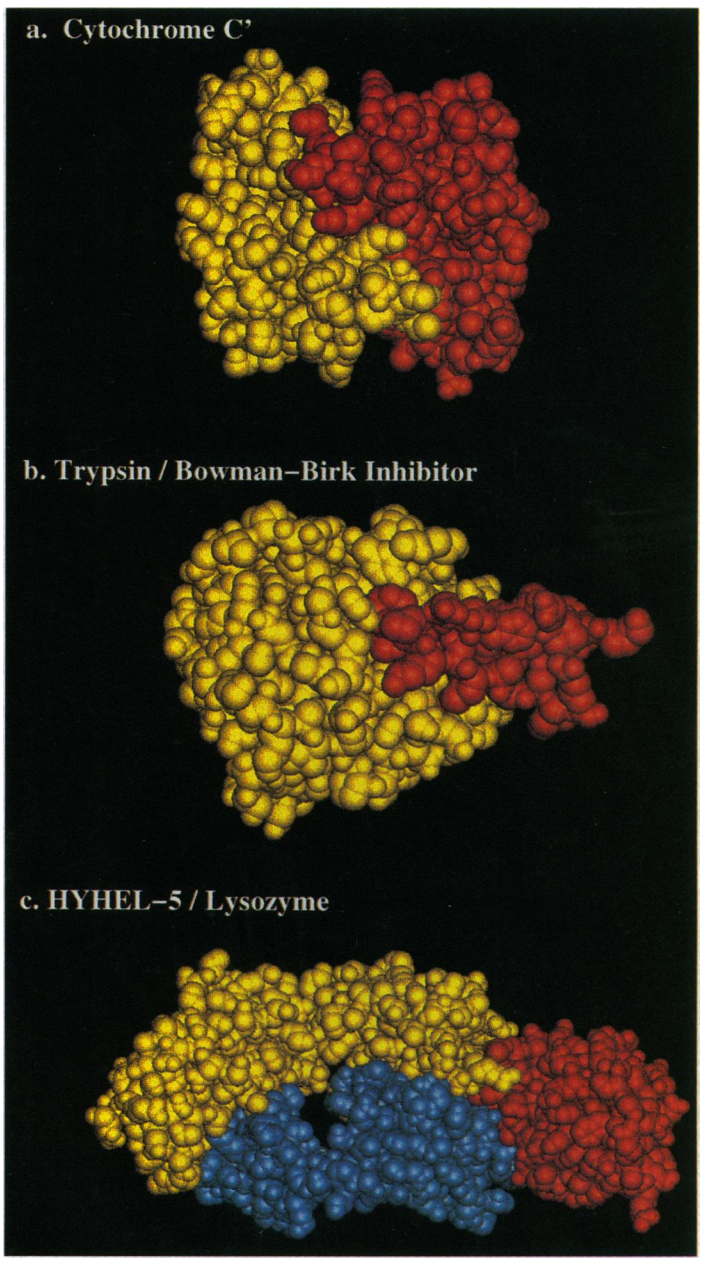

FIG. 1. Corey–Pauling–Koltun models of protein–protein complexes. The complex components have been differentiated by color, and it should be noted that the scales are not comparable between the different structures. (*a*) Homodimer: cytochrome *c'* (PDB code 2ccy) (1). Subunit A is in yellow and subunit B is in red. (*b*) Enzyme–inhibitor complex: trypsin and inhibitor from bitter gourd (PDB code 1tab) (2). The enzyme is in yellow and the much smaller inhibitor is in red. (*c*) Antibody–protein complex: HYHEL-5–lysozyme (PDB code 2hfl) (3). The light and heavy chains of the Fab are colored yellow and blue and the lysozyme is in red.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

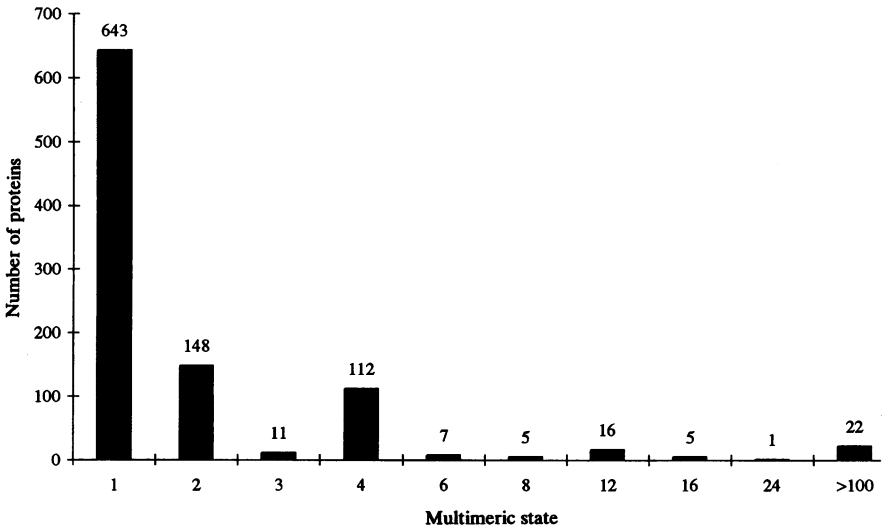

FIG. 2. Multimeric states of proteins in the July 1993 PDB (4). 1 = monomer, 2 = dimer, etc.

(11), and Jones and Thornton (12) have all compared structural properties (including hydrophobicity, accessible surface area, shape, and residue preferences) between interior, surface, and interface components in oligomeric proteins. The comparison of different types of complexes (enzyme–inhibitor and antibody–antigen) in terms of interface size and hydrophobicity has also been addressed (13, 14). More recent work has centered on the prediction of interface sites using residue hydrophobicity. Korn and Burnett (15) used hydropathy analysis to predict the position of the interface in a dimeric protein using a nonautomated method. Young *et al.* (16) have taken this approach further and produced an automated predictive algorithm based on the analysis of the hydrophobicity of clusters of residues in proteins.

In this review, we study 59 different complexes found in the PDB (4), which can be divided into four different types (Table 1).

(*i*) Thirty-two nonhomologous homodimers: proteins with two identical subunits.

(*ii*) Ten enzyme–inhibitor complexes.

(*iii*) Six antibody–protein complexes.

(*iv*) Eleven "other" heterocomplexes including 4 permanent complexes and 7 interfaces between independent monomers.

These different types of complexes have different biological roles. Most homodimers are only observed in the multimeric state, and it is often impossible to separate them without denaturing the individual monomer structures. Many homodimers also have twofold symmetry, which places additional constraints on their intersubunit relationship. Many enzyme–inhibitor complexes are also strongly associated, with binding constants ranging from 10^{-7} mol^{-1} to 10^{-13} mol^{-1} (17), yet these molecules also exist independently as stable entities in solution. Similarly, the antibody–protein complexes and six of the other heterocomplexes are composed of molecules that have an independent existence.

From the evolutionary perspective, the homodimers, enzyme–inhibitors, and the heterocomplexes have presumably all evolved over time to optimize the interface to suit their biological function. In some examples, the function may require the evolution of strong binding, while other circumstances may dictate weaker binding. In contrast, the antibody–protein interactions are relatively "happenstance" and are selected principally by the strength of the binding constant, without being subject to evolutionary optimization over many years. Thus in the following review, we attempt to characterize the interactions observed between proteins in the light of their biological function.

For the current work, protein–protein interfaces have been defined based on the change in their solvent accessible surface area (ΔASA) when going from a monomeric to a dimeric state. The ASAs of the complexes were calculated using an implementation of the Lee and Richards (20) algorithm developed by Hubbard (21). The interface residues (atoms) were defined as those having ASAs that decreased by >1 Å^2 (0.01 Å^2) on complexation.

2. Characterization of Protein–Protein Interfaces

There are several fundamental properties that characterize a protein–protein interface, which can be calculated from the coordinates of the complex.

2.1. Size and Shape. The size and shape of protein interfaces can be measured simply in absolute dimensions (Å) or, more accurately, in terms of the ΔASA on complexation. The ΔASA method was used, as it is known that there is a correlation between the hydrophobic free energy of transfer from polar to a hydrophobic environment and the solvent ASA (22). Thus, calculating ΔASA may provide a measure of the binding strength. The shape of the interfaces is also analyzed, as this is relevant to designing molecular mimics.

The mean ΔASA on complexation (going from a monomeric state to a dimeric state) was calculated as half the sum of the total ΔASA for both molecules for each type of complex (Table 2). To give a guide of how much of a protein subunit's surface is buried on complexation, the ΔASA values for individual complexes were compared with the molecular weights of the constituent subunits (Fig. 3). For the heterocomplexes, the molecular weights will be different for each component and hence the smaller component was used, as this will limit the maximum size of the interface.

In the homodimers, the ΔASA varies widely from 368 Å^2 to 4746 Å^2, and there is a clear, though scattered, relationship with the molecular weight of the subunit [correlation coefficient (r) is 0.69], with the larger molecules in general having larger interfaces. The range of ΔASA in the heterocomplexes is smaller (639 Å^2 to 3228 Å^2). This constancy presumably reflects three factors: the limited nature of the PDB, the average size of protein domains, and the biological constraints. In addition, it should be noted that all the enzyme–inhibitor complexes involve proteinases, and, with the exception of papain and subtilisin, all are related to trypsin, although the corresponding inhibitors are nonhomologous.

In Fig. 3*b* it can be seen that three heterocomplexes [cathespin D (PDB code 1lya), reverse transcriptase (PDB code 3hvt), and human chorionic gonadotropin (PDB code 1hrp)] have relatively large interfaces for their molecular weights These three are all permanent complexes, and the size of the interfaces in these structures is more comparable with the distribution observed in the homodimers (Fig. 3*a*) than the heterocomplexes.

Two protein subunits may interact and form a protein–protein interface with two relatively flat surfaces or form a twisted interface. To assess how flat or how twisted the protein–protein interfaces were, a measure of how far the interface residues deviated from a plane (termed planarity) was calculated. The planarity of the surfaces between two components of a complex was analyzed by calculating the rms deviation of all the interface atoms from the least-squares plane through the atoms. Fig. 4 shows that the heterocomplexes have interfaces that are more planar than the homodimers. The higher mean rms deviation of the homodimers

Abbreviations: PDB, Brookhaven Protein Databank; ASA, accessible surface area; ΔASA, change in the accessible surface area; HIV, human immunodeficiency virus.

Table 1. Data sets of protein–protein complexes

PDB code	Protein	Resolution, Å
	*Nonhomologous homodimers**	
1cdt	Cardiotoxin	2.5
1fc1	Fc fragment (immunoglobulin)	2.9
1il8	Interleukin	NMR
1msb	Mannose binding protein	2.3
1phh	*p*-Hydroxybenzoate hydrolase	2.3
1pp2	Phospholipase	2.5
1pyp	Inorganic pyrophosphatase	3.0
1sdh	Hemoglobin (clam)	2.4
1utg	Uteroglobin	1.35
1vsg	Variant surface glycoprotein	2.9
1ypi	Triose phosphate isomerase	1.9
2ccy	Cytochrome *c*3	1.67
2cts	Citrate synthase *c*	2.0
2gn5	Gene 5 DNA-binding protein	2.3
2or1	434 repressor	2.5
2rhe	Bence–Jones protein	1.6
2rus	Rubisco	2.3
2rve	*Eco*RV endonuclease	3.0
2sod	Superoxide dismutase	2.0
2ssi	Subtilisin inhibitor	2.6
2ts1	Tyrosyl transferase RNA synthase	2.3
2tsc	Thymidylate synthase	1.97
2wrp	Trp repressor	1.65
3aat	Aspartate aminotransferase	2.8
3enl	Enolase	2.25
3gap	Catabolite gene activator protein	2.5
3grs	Glutathione reductase	1.54
3ied	Isocitrate dehydrogenase	2.5
3sdp	Iron superoxidase	2.1
4mdh	Cytoplasmic malate dehydrogenase	2.5
5adh	Alcohol dehydrogenase	2.9
5hvp	HIV protease	2.0
	Enzyme–inhibitor complexes†	
1ach	α-Chymotrypsin–eglin C	2.0
1cho	α-Chymotrypsin–ovomucoid third domain	1.8
1cse	Subtilisin Carlsberg–eglin C	1.2
1mct	Trypsin–inhibitor from bitter gourd	1.6
1mcc	Peptidyl peptide hydrolase–Eglin C	2.0
1stf	Papain–inhibitor stefin B mutant	2.37
1tab	Trypsin–Bowman–Birk inhibitor	2.3
1tgs	Trypsinogen–Pancreatic secretory trypsin inhibitor	1.8
2ptc	β-Trypsin–pancreatic trypsin inhibitor	1.9
2sic	Subtilisin–streptomyces subtilisin inhibitor	1.8
	Antibody–antigen complexes‡	
1fdl	D1.3 Fab–hen egg white lysozyme	2.5
1jel	Fab JE142–histidine containing protein	2.8
1jhl	D11.15 Fv–pheasant egg lysozyme	2.4
1nca	NC41 Fab/influenza virus N9 neuraminidase	2.5
2hfl	HYHEL-5 Fab–chicken-lysozyme	2.54
3hfm	HYHEL-10 Fab–chicken lysozyme	3.0
	Other heterodimeric complexes§	
1atn	Deoyribonuclease I–actin	2.8
1gln	Glycerol kinase–glucose-specific factor III	2.6
1hrp¶	Human chorionic gonadotropin	3.0
1lpa	Lipase–colipase	3.04
1lya¶	Cathepsin D	2.5
2btf	β-Actin–profilin	2.55
2pch	Yeast cytochrome *c* peroxidase–horse cytochrome *c*	2.8
3hhr‖	Human growth hormone–human growth hormone receptor	2.8
3hvt¶	Reverse transcriptase	2.9
6rlx¶,**	Relaxin	1.5

*Data set of 32 nonhomologous homodimers. Protein dimers were selected for inclusion on the basis that they had a sequence identity of <35% and were structurally different. The structural similarity of the proteins was measured using a method of direct structural alignment, SSAP (18). Proteins were selected for the data set if they had a SSAP score of ≤80. In the process of selection, only dimers with identical subunits were considered. This selection resulted in a nonhomologous data set of 32 protein dimers, each belonging to a different homologous protein family.

†Data set of 10 enzyme–inhibitor complexes. These heterocomplexes were selected from the PDB such that, although the enzymes components could be homologous, the corresponding inhibitors were nonhomologous (for definition of nonhomologous see footnote *) or vice versa.

‡Data set of six antibody–protein complexes. Although homologous pairs are included (e.g., four antibody–lysozyme complexes), the sites of recognition on the lysozyme are different.

§Data set of other heterocomplexes. This data set contains those complexes selected from the PDB which did not fit into either of the other heterocomplex categories. This data set includes four structures that occur only in the complexed form and six structures that occur as both complexes and as monomers.

¶Occur only as heterodimers.

‖Contributes two hormone–receptor complexes.

**Derived from a single chain precursor.

results from five proteins that had comparatively high rms deviation values (>6 Å). These are dimers in which the two subunits were twisted together across the interface [e.g., isocitrate dehydrogenase (26)] or proteins that had subunits with "arms" apparently clasping the two halves of the structure together [e.g., aspartate aminotransferase (28)] (Fig. 5). When the other heterocomplex data set is divided into structures that occur only as heterodimers and those that occur as both heterocomplexes and monomers (Table 2), it becomes apparent that the former resemble the homodimers in that they are less planar compared to their nonpermanent counterparts, which occur as both monomers and as dimer complexes.

To provide a rough guide to the shape of the interface, the "circularity" of the interfaces was calculated as the ratio of the lengths of the principal axes of the least-squares plane through the atoms in the interface. A ratio of 1.0 indicates that an interface is approximately circular. The shape of the interface region (Table 2) varies little between the homodimers, the antigens, and the enzyme component of the enzyme–inhibitor complex; each type is relatively circular with an average ratio of between 0.71 and 0.75. In comparison, the inhibitors of the enzyme–inhibitor complexes have less circular interfaces, with an average ratio of 0.55. The homodimers show the largest variation, with the elongated interface of variant surface glycoprotein of *Trypanosoma brucei* (29) at one extreme (ratio 0.25). The interface of this structure, which forms a coat on the surface of the parasite, reflects the elongated nature of the protein as a whole.

2.2. Complementarity Between Surfaces. Many authors have commented on the electrostatic and the shape complementarity observed between associating molecules (5, 30–33). The electrostatic complementarity between interfaces has been used as an additional filter for many protein–protein docking methods (see, for example, ref. 34) and new methods of evaluating shape complementarity have been evolved (see, for example, ref. 33).

In this review, the complementarity of the interacting surfaces in the protein–protein complexes has been evaluated by defining a gap index:

$$\text{Gap index (Å)} = \text{gap volume between molecules (Å}^3\text{)} / \text{interface ASA (Å}^2\text{) (per complex).} \quad [1]$$

The gap volume was calculated using a procedure developed by Laskowski (23), which estimates the volume enclosed between any two molecules, delimiting the boundary by defining a maximum allowed distance from both interfaces. A mean gap index was calculated for each type of complex (Table 2). The results indicate that the interacting surfaces in the ho-

Table 2. Results of structural analysis on protein–protein complexes

Characteristic	Homodimer	All hetero-complexes	Enzyme–inhibitor		Antibody–protein	Other hetero-complexes	
No. of examples	32	27	10		6	7*	4†
ΔASA,‡ Å²							
Mean	1685.03	983.06	785.10		777.42	848.67	2021.59
σ	1101.09	582.03	74.52		135.33	243.63	1036.64
Planarity,§ Å							
Mean	3.46	2.80	2.70		2.21	2.54	3.94
σ	1.72	0.87	0.45		0.39	0.61	1.35
Circularity¶			E	I	Ag	BS	BS
Mean	0.71		0.73	0.55	0.75	0.61	0.64
σ	0.17		0.05	0.10	0.12	0.16	0.23
Minimum	0.25		0.62	0.43	0.62	0.41	0.30
Maximum	1.00		0.78	0.71	0.92	0.91	0.93
Segmentation‖			E	I	Ag	BS	BS
Mean	5.22		7.8	2.7	3.83	4.64	5.63
σ	2.55		1.03	0.95	1.83	1.28	3.85
Minimum	2		6	2	2	3	1
Maximum	11		9	5	7	7	11
Hydrogen bonds per 100 A² ΔASA**							
Mean	0.70	1.13	1.37		1.06	0.85	1.10
σ	0.46	0.47	0.37		0.51	0.37	0.61
Minimum	0	0.29	0.60		0.47	0.39	0.29
Maximum	1.7	1.88	1.87		1.88	1.47	1.65
Gap index,†† Å							
Mean	2.20	2.48	2.20		3.02	3.02	1.47
σ	0.87	1.02	0.47		0.80	1.13	1.34
Minimum	0.57	0.35	1.38		2.04	2.03	0.35
Maximum	4.43	5.17	2.86		3.96	5.17	3.42
Hydrophobicity‡‡							
Mean interior	+0.26	+0.26	+0.25		+0.28	+0.29	+0.19
Mean interface	+0.12	−0.14	−0.03		−0.22	−0.26	−0.05
Mean exterior	−0.27	−0.23	−0.21		−0.24	−0.27	−0.19

E, enzyme; I, inhibitor; Ag, antigen; BS, both subunits.

*Occur as monomers and as heterodimer complexes.

†Occur only as heterodimers.

‡ΔASA for one subunit on complexation. In the heterodimer data sets, enzyme–inhibitor and other heterocomplexes, the ASA shown is the mean ASA buried by each subunit. In the antibody–protein data set, the ASA is the ASA on the antigen (protein) surface buried on complexation.

§A program [implemented by Laskowski (23)] was used to calculate the atomic rms deviation of all the interface atoms from the least-squares plane through all these atoms.

¶The circularity of the interfaces was calculated as the ratio of the lengths of the principal axises of the least-squares plane through the atoms in the interface. The standard deviation (σ) and range of the distribution are also shown.

‖It was defined that interface residues separated by more than five residues were allocated to different segments. The standard deviation (σ) and the range of the distribution are also shown.

**The number of intermolecular hydrogen bonds per 100 Å² ΔASA were calculated using HBPLUS (24) in which hydrogen bonds are defined according to standard geometric criteria. The standard deviation (σ) and range of the distribution are also shown.

††The gap volumes between the two components of the complexes was calculated using SURFNET (18).

‡‡Mean hydrophobicity values [derived using the scale of Janin *et al.* (9) based on statistical analysis of protein structures] for three subsets of residues within each type of protein–protein complex. The interface residues were defined as explained in section 2. The definition of exterior and interior residues were based on relative ASA of each residue, which range from 0% for residues with no atom contact with the solvent to 100% for fully accessible residues. On this basis, the exterior residues were defined as having relative accessibilities >5%, and interior residues were defined as those with relative accessibilities ≤5%. This 5% cut-off was devised and optimized by Miller *et al.* (47). The subset of interfaces residues were excluded from the subsets of interiors and exteriors, resulting in three discrete sets of residues for each of the complexes.

modimers, the enzyme–inhibitor complexes, and the permanent heterocomplexes (a subset of the other heterocomplex data set) are the most complementary, whereas the antibody–antigen complexes and the nonobligatory other heterocomplexes are the least complementary (although all four distributions do overlap considerably). These data agree with the conclusions drawn by Lawrence and Colman (33), using their shape complementarity statistic.

2.3. Residue Interface Propensities. The relative importance of different amino acids residues in the interfaces of complexes can give a general indication of the hydrophobicity. Such information can only be interpreted if the distribution of residues occurring in the interface are compared with the distribution of residues occurring on the protein surface as a whole. Residue interface propensities were calculated for each amino acid (AA_j) as the fraction of ASA that AA_j contributed to the interface compared with the fraction of ASA that AA_j contributed to the whole surface (exterior residues plus interface residues) i.e.,

Interface residue propensity AA_j =

$$\left(\sum_{i=1}^{Ni} \mathrm{ASA}_{\mathrm{AA}_j(i)} \Big/ \sum_{i=1}^{Ni} \mathrm{ASA}_{(i)}\right) \Big/ \left(\sum_{i=1}^{Ns} \mathrm{ASA}_{\mathrm{AA}_j(s)} \Big/ \sum_{i=1}^{Ns} \mathrm{ASA}_{(s)}\right), \quad [2]$$

where $\Sigma\mathrm{ASA}_{\mathrm{AA}_j(i)}$ = sum of the ASA (in the monomer) of amino acid residues of type j in the interface, ΣASA(i) = sum of the ASA (in the monomer) of all amino acid residues of all types in the interface, $\Sigma\mathrm{ASA}_{\mathrm{AA}_j(s)}$ = sum of the ASA (in the monomer) of amino acid residues of type j on the surface (exterior plus interface residues), and ΣASA(s) = sum of the ASA (in the monomer) of all amino acid residues of all types on the surface.

A propensity of >1 denotes that a residue occurs more frequently in the interface than on the protein surface. The propensities (Fig. 6) show that, with the exception of methionine, the hydrophobic residues show a greater preference for the interfaces of homodimers than for those of heterocomplexes. The lower propensities for hydrophobic residues in the heterocomplex interfaces is balanced by an increased propensity for the polar residues.

2.4. Hydrophobicity Including Hydrogen Bonding. It has often been assumed that proteins will associate through hydrophobic patches on their surfaces. However, polar interactions between subunits are also common and, in terms of the driving force for complexation, it is important to explore the relative contributions of these effects, including reference to the subunits' ability to exist independently.

A mean hydrophobicity value [based on the scale derived by Janin *et al.* (9)] was calculated for all residues defined in the interface of each complex. A mean value was calculated for each type of complex and for all heterocomplexes (Table 2). In all of the complexes, the interface has an intermediate hydrophobicity between those of the interior (hydrophobic) and the exterior (hydrophilic). When the hydrophobicity values of the interface are compared between the homodimers and the heterocomplexes, it is seen that, as previously concluded from the residue propensities, the

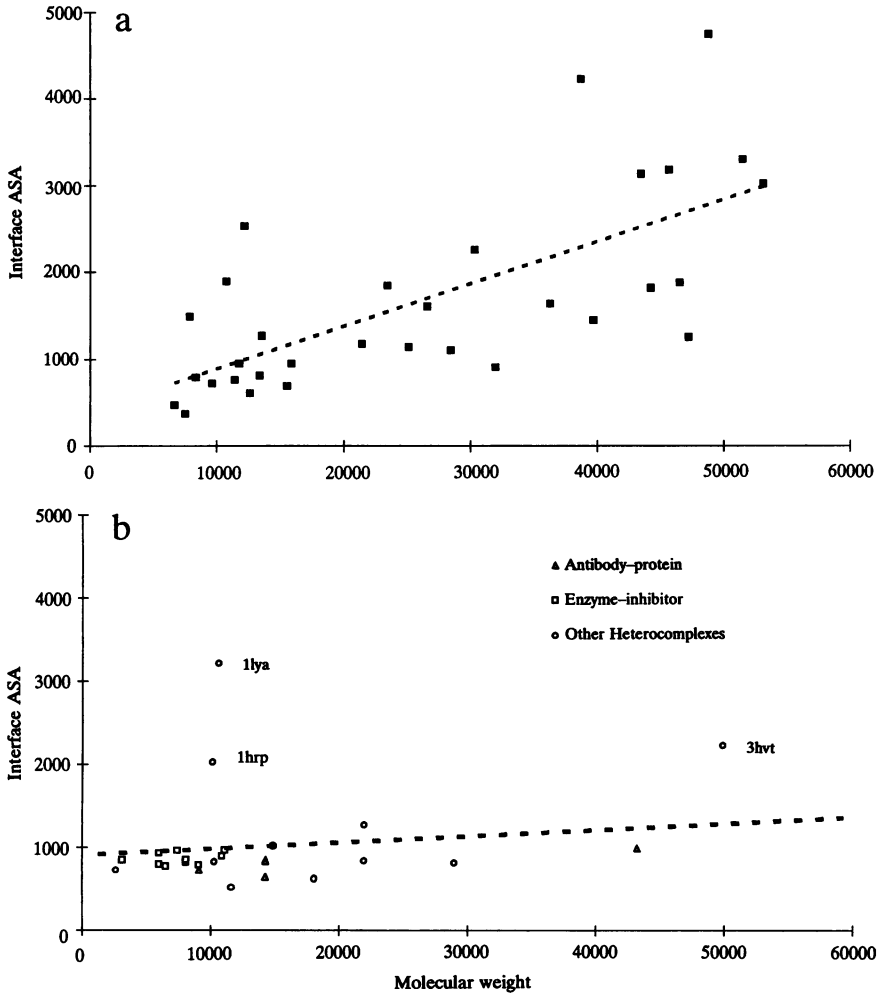

FIG. 3. (*a*) Interface ASA vs. molecular weight for homodimers. The ASA (measured in Å^2) is that buried by one subunit on dimerization, and the molecular weight is that of the monomer. The dashed line is the straight line regression (r = 0.69). (*b*) Interface ASA vs. molecular weight for heterocomplexes. The ΔASA (measured in Å^2) and the molecular weight are both from the smallest subunit. The dashed line is the straight line regression for all heterocomplexes (r = 0.17).

interfaces of the heterocomplexes are less hydrophobic than those of homodimers. This difference in hydrophobicity, which has previously been observed (9, 13), can be explained by the roles of the two types of complex. The homodimers rarely occur or function as monomers, and hence their hydrophobic surfaces are permanently buried within a protein–protein complex. Of the 27 heterocomplexes analyzed in this review, 23 (10 enzyme–inhibitor complexes, 6 antibody–protein complexes, and 7 other heterocomplexes) do occur as monomers in solution and have biological functions in this state. Hence these interfaces cannot be as hydrophobic as those of the homodimers, because a large exposed hydrophobic patch on the protein would be energetically unfavorable.

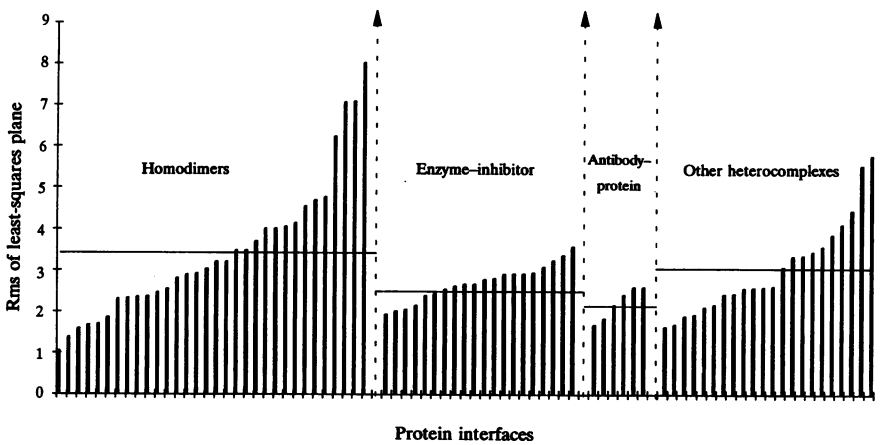

FIG. 4. Planarity of protein–protein interfaces. The rms deviation of atoms from the least-squares plane through these atoms is shown for one subunit of the homodimers, for both subunits of the enzyme–inhibitor complexes and other heterocomplexes, and for the antigen subunit of the antibody–antigen complexes. Within each group of complexes, the proteins have been placed in ascending order of rms deviation. The mean of each data set is indicated by a solid horizontal line. Each bar represents one interface for a single protein.

To identify the major polar interactions between the components in the complexes, the mean number of hydrogen bonds per 100 Å^2 of ΔASA was calculated for each type of complex (Table 2). The 23 heterocomplexes that occur as both monomers and complexes have relatively more intermolecular hydrogen bonds per ΔASA. The four heterocomplexes that occur only as heterodimers show a similar numbers of hydrogen bonds per 100 Å^2 of ΔASA as the homodimers. This distribution was expected from the residue propensities, which showed that the transient complexes (those with components that occur as both monomers and complexes) contained more hydrophilic residues in their interfaces than the permanent complexes.

2.5. Segmentation and Secondary Structure. The number of discontinuous segments of the polypeptide chain involved in the interface is important since the ability of peptides or small molecules to mimic one-half of the interaction may depend upon it. For example, in an interface that is dominated by one segment, a single peptide will probably be a good mimic. However, the design of molecules to mimic multisegmented interfaces will almost certainly be more difficult.

To analyze the discontinuous nature of the interfaces, in terms of the amino acid sequence, the mean number of segments in the interfaces was calculated for each type of complex (Table 2). It was defined that interface residues separated by more than 5 residues were allocated to different segments. In the 59 complexes studied, the number of segments varies from 1 to 11. In fact, only 1 complex [relaxin (35)] had one segment at the interface, as it is a very small protein derived from a single chain precursor, with only 24 residues in the α-chain and 29 in the β-chain. The enzyme–inhibitor complexes are unusual in having only two to five segments interacting. This class of inhibitors has evolved to mimic an elongated segment of polypeptide chain, in the conformation required for cleavage by the enzyme, and therefore all present a protruding canonical loop structure (36, 37), which dominates the interaction. In contrast the other interfaces are highly segmented, especially the long binding site cleft in the proteinases, which on average contains seven segments.

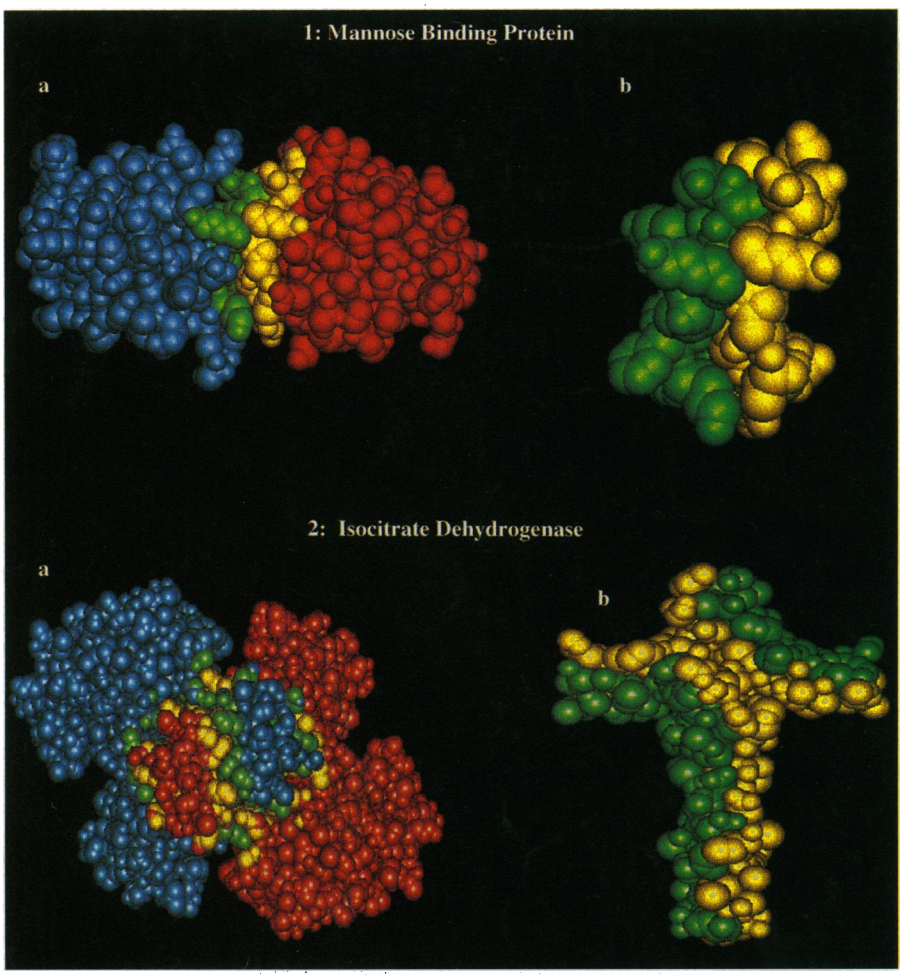

FIG. 5. Corey–Pauling–Koltun models of planar and nonplanar interfaces in protein complexes. (*Upper*) Two subunits are shown: one subunit is colored blue and one red. The interface atoms in each subunit are colored differently; the atoms in green are the interface atoms in the blue subunit and those in yellow are the interface atoms in the red subunit. (*Lower*) Only the interface atoms of the two structures are shown. (*Upper*) Mannose binding protein (PDB code 1msb) (27): a planar interface. (*a*) Dimer viewed looking along the subunit interface. (*b*) Dimer interface only shown. (*Lower*) Isocitrate dehydrogenase (PDB code 3icd) (26): a nonplanar interface. (*a*) Dimer viewed looking down the subunit interface showing the two subunits twisted together at the top. (*b*) Dimer interface only shown, viewed along the interface.

The secondary structure of the interface regions has also been analyzed. Over the whole data set, it was found that there is an approximately equal proportion of helical, strand, and coil residues involved. Some interfaces contain only one type of structure (helices, strands, or loops), but most are mixed. The interfaces involving β sheets fall into three categories, those that interact by extending the sheet through classic mainchain hydrogen bonding [e.g., human immunodeficiency virus (HIV) protease (38)], those in which the sheets stack on top of one another [e.g., subtilisin inhibitor homodimer (39)], and mixed structures in which the β sheets are neither clearly stacked nor extended [e.g., copper, zinc, and superoxide dismutase (40)].

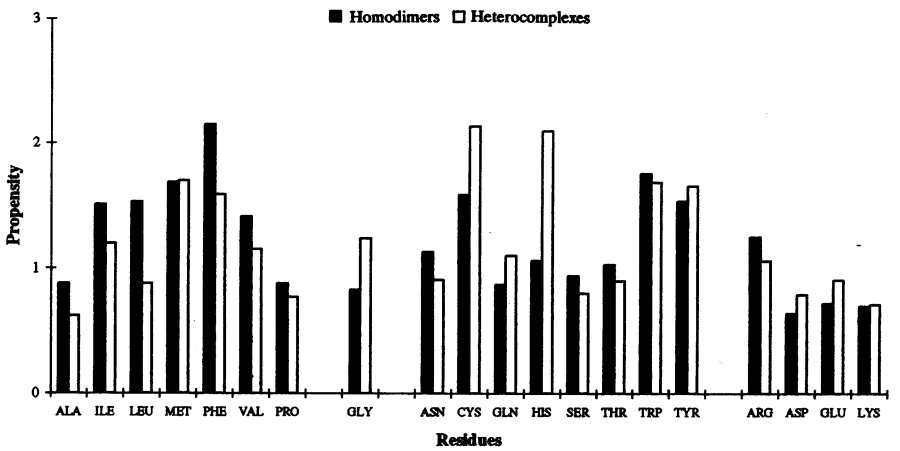

FIG. 6. Residue interface propensities were calculated for each amino acid (AA_j) based on the fraction of ASA that AA_j contributed to the interface compared with the fraction of ASA that AA_j contributed to the whole surface (exterior residues plus interface residues) (see section 2.2).

2.6. Conformational Changes on Complex Formation. It is not clear to what extent proteins change their conformation on forming a complex (36, 41, 42), and currently there are few proteins that have been structurally determined (by crystallography or nuclear magnetic resonance) before and after complexation. However, it is possible to distinguish various levels of conformational change: no change, side chain movements alone, segment movement involving the mainchain (e.g., hinged loop), and domain movements (gross relative movements of the domains). The mechanism of domain movement is specifically relevant to enzyme complexes, which often undergo domain shifts when binding substrates [e.g., adenylate kinase (43) and lactoferrin (43)]. For antibody–protein recognition, there is a wide range of variation that can occur on binding (41, 42, 45, 46). Overall we can expect that both rigid and flexible docking will occur in different circumstances, but there will always be an energetic price to pay for reducing flexibility.

3. Patch Analysis of Protein Surfaces in Homodimers

So far we have analyzed the interface regions in isolation, but it is also instructive to explore whether these regions are significantly different from the rest of the protein surface in any way. The problem to be addressed is *given a protein of known structure (but with no known structure for its complex) is it possible to identify the interface region on its surface?* Here we use the monomer structures of the homodimers and compare their surface residue patches.

A patch is defined as a central *surface-accessible* residue with n nearest *surface-accessible* neighbors, as defined by C^{α} positions, where n is taken as the number of residues observed in the known homodimer interface. A number of constraints was used to ensure that the residues selected in a patch represented a contiguous patch on the surface of the protein.

This procedure defines a number of overlapping patches of accessible residues. For example, in the HIV protease structure (PDB code 5hvp) (38), there are 81 such patches. Each possible surface patch was then analyzed for a series of parameters including, residue propensity (section 2.3). ASA, protrusion index (44), planarity (section 2.1), and hydrophobic-

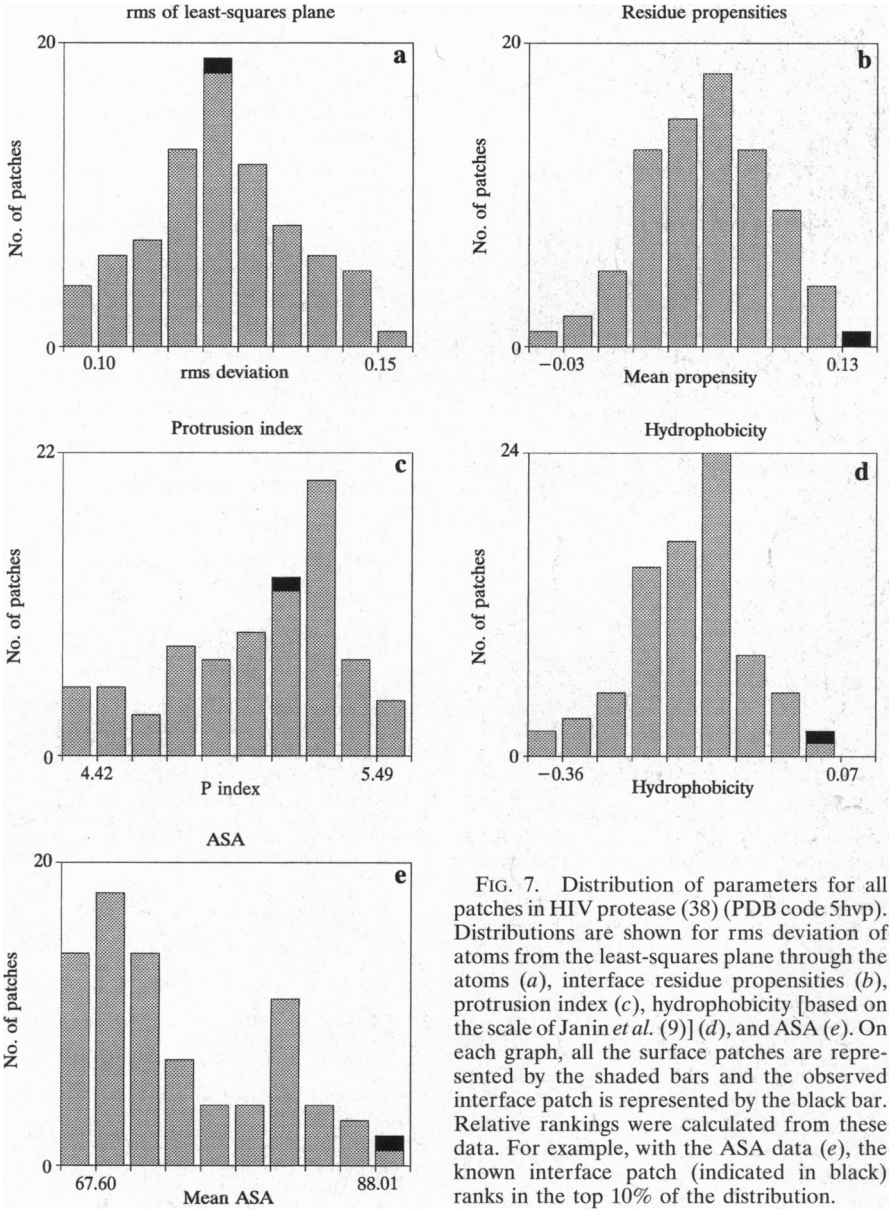

FIG. 7. Distribution of parameters for all patches in HIV protease (38) (PDB code 5hvp). Distributions are shown for rms deviation of atoms from the least-squares plane through the atoms (*a*), interface residue propensities (*b*), protrusion index (*c*), hydrophobicity [based on the scale of Janin *et al.* (9)] (*d*), and ASA (*e*). On each graph, all the surface patches are represented by the shaded bars and the observed interface patch is represented by the black bar. Relative rankings were calculated from these data. For example, with the ASA data (*e*), the known interface patch (indicated in black) ranks in the top 10% of the distribution.

ity (section 2.4). These parameters were also evaluated for the known residue interfaces. Thus, for each parameter the distribution of values for all the patches on one protein, including the observed interface patch, can be plotted [for example HIV protease (PDB code 5hvp) (38); Fig. 7]. A ranking of the true interface patch relative to the other possible patches (e.g., top 10%, 10–20%, etc.) was then calculated. With this approach, it becomes possible to plot the rankings of all the observed patches for each protein as a histogram (Fig. 8) to assess which parameters best differentiate the interface region. The aim is to identify likely recognition sites from a structure for which structural data on the complex is not available.

It can be seen that no single parameter absolutely differentiates the interfaces from all other surface patches. For example, with the planarity parameter, 50% of the interfaces were in the most planar bin (i.e., among the top 10% of patches that were most planar), but others were very nonplanar (see section 2.1). The most striking correlation is for the accessible surface area (Fig. 8*e*). This observation in part reflects the fact that the side chains from one monomer extend from the surface to interact with the other half of the dimer. In isolation, therefore, they become highly accessible, and we would not expect to see such a strong signal for the structure of an isolated molecule prior to complexation, as the side chains probably change their conformation and "stretch out" to form the complex. As expected from the accessibility data, the interfaces tend to protrude from the surface (Fig. 8*c*), although the signal is weaker, perhaps as a consequence of the requirements for planarity. Of course some recognition regions are more concave (e.g., the antibody-combining site), but for the homodimers the general trend is to favor protrusion. Similarly the residue propensities (Fig. 8*b*) show some discriminating power, suggesting that the index does carry relevant information, although the trend is not as marked as for some of other of the parameters. The weakest correlation can be seen for the "hydrophobicity" measure (Fig. 8*d*) derived from the Janin *et al.* (9) parameters, although even here there is some suggestion that the interface patch tends toward the hydrophobic.

None of the distributions are definitive in that their interface region is never always at one extreme, but they all show trends for the known interface to be distinguished from other surface patches. This type of comparative analysis, including many different parameters rather than a single value, can potentially be used to predict the location of likely interface sites on protein surfaces.

For a protein that is known to be involved in protein–protein interactions and whose structure has been determined but for which there is no structure for the complex available, it is straightforward to analyze the surface patches and calculate their properties as shown in Fig. 8. For each patch we can calculate a combined probability that it will be involved in forming an interface to another protein molecule. These probabilities can be rank-ordered to identify putative interfaces. Using this method for the homodimers, we can identify >70% of the interface regions correctly. Such an approach is useful for identifying candidate interface residues, which can be mutated experimentally and tested for the effect on complex formation.

4. Discussion

This review has highlighted the need to take into account the type of protein–protein complexes (as shown in Table 1) when characterizing the interfaces within them. Complexes can be permanent or nonobligatory. The requirement for the molecules to exist as independent entities imposes additional constraints on these structures, and their interfaces are less hydrophobic than those that only exist in a multimeric form. In addition, it was found that the permanent complexes had protein–protein interfaces that were more closely packed but less planar and with fewer intersubunit hydrogen bonds than the nonobligatory complexes.

The results presented here are derived from a relatively small data set of protein complexes. This analysis has been difficult because of the lack of information on the *in vivo* complex status in the current PDB entries, so that extracting all dimers, for example, is a very labor-intensive process. It is also important to recognize related complexes, so that a data set is not biased. Clearly this work needs to be extended. As the data base grows rapidly, we would like to include higher order complexes and such

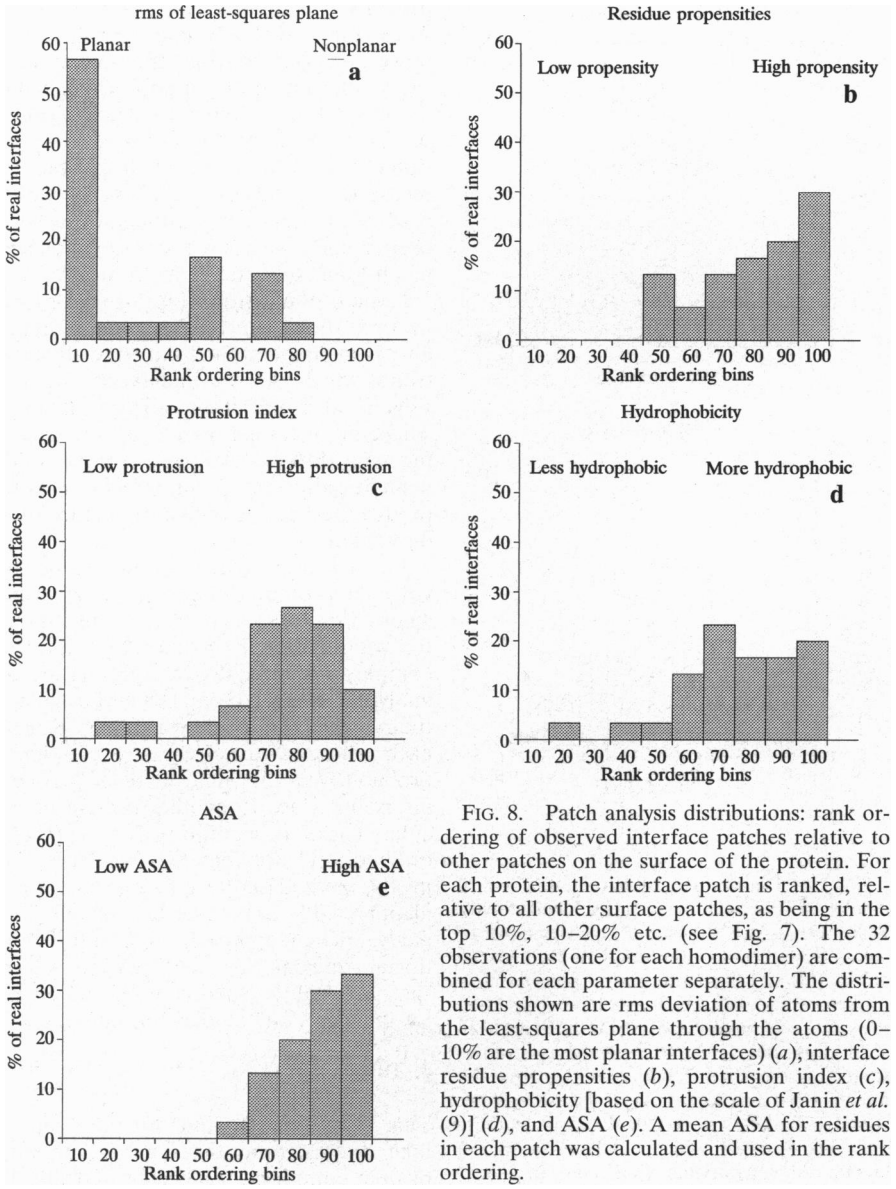

FIG. 8. Patch analysis distributions: rank ordering of observed interface patches relative to other patches on the surface of the protein. For each protein, the interface patch is ranked, relative to all other surface patches, as being in the top 10%, 10–20% etc. (see Fig. 7). The 32 observations (one for each homodimer) are combined for each parameter separately. The distributions shown are rms deviation of atoms from the least-squares plane through the atoms (0–10% are the most planar interfaces) (*a*), interface residue propensities (*b*), protrusion index (*c*), hydrophobicity [based on the scale of Janin *et al.* (9)] (*d*), and ASA (*e*). A mean ASA for residues in each patch was calculated and used in the rank ordering.

factors as interdigitation, conformational change on complex formation, and correlation with binding constants. The latter are often difficult to determine experimentally and are almost never deposited with the coordinates, yet they are essential if we are to understand the kinetics and thermodynamics of complex formation.

What is clear is that over the next few years, there will be a cascade of coordinate data for protein–protein interactions. We will almost certainly see more nonobligatory complexes, with weaker interactions, as these are often of great biological relevance. In nature many of the most important biological functions involve huge multicomponent complexes (e.g., the ribosome), and we are only just taking our first steps to understand the principles of molecular recognition in simple systems. However, the implications of a better understanding for the design of new therapeutics and environmental products are apparent to all. The next few years promise much excitement as we discover more about how proteins interact together to perform their biological function.

S.J. is funded by the Biotechnology and Biological Research Council and Zeneca Pharmaceuticals.

1. Finzel, B. C., Weber, P. C., Hardman, K. D. & Salemme, F. R. (1985) *J. Mol. Biol.* **186,** 627–643.
2. Tsunogae, Y., Tanaka, I., Yamane, T., Kikkawa, J. I., Ashida, T., Ishikawa, C., Watanabe, K., Nakamura, S. & Takahashi, K. (1986) *J. Biochem. (Tokyo)* **100,** 1637–1645.
3. Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 8075–8092.
4. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
5. Johnson, J. E. (1996) *Proc. Natl. Acad. Sci. USA* **92,** 27–33.
6. Walls, P. H. & Sternberg, M. J. E. (1992) *J. Mol. Biol.* **228,** 277–297.
7. Helmer-Citterich, M. & Tramontano, A. (1994) *J. Mol. Biol.* **324,** 1021–1031.
8. Zielenkiewicz, P. & Rabczenko, A. (1988) *Biophys. Chem.* **29,** 219–224.
9. Janin, J., Miller, S. & Chothia, C. (1988) *J. Mol. Biol.* **204,** 155–164.
10. Miller, S. (1989) *Protein Eng.* **3,** 77–83.
11. Argos, P. (1988) *Protein Eng.* **2,** 101–113.
12. Jones, S. & Thornton, J. M. (1995) *Prog. Biophys. Mol. Biol.* **63,** 31–65.
13. Janin, J. & Chothia, C. (1990) *J. Biol. Chem.* **265,** 16027–16030.
14. Duquerroy, S., Cherfils, J. & Janin, J. (1991) *Ciba Found. Symp.* **161,** 237–252.
15. Korn, A. P. & Burnett, R. M. (1991) *Proteins: Struct. Funct. Genet.* **9,** 37–55.
16. Young, L., Jernigan, R. L. & Covell, D. G. (1994) *Protein Sci.* **3,** 717–729.
17. Laskowski, M. & Kato, I. (1980) *Annu. Rev. Biochem.* **49,** 593–626.
18. Taylor, W. R. & Orengo, C. A. (1989) *J. Mol. Biol.* **208,** 1–22.
19. Wells, J. A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1–6.
20. Lee, B. & Richards, F. M. (1971) *J. Mol. Biol.* **55,** 379–400.
21. Hubbard, S. J. (1992) PhD. thesis (Univ. of London, London, England).
22. Chothia, C. (1974) *Nature (London)* **248,** 338–339.
23. Laskowski, R. A. (1991) SURFNET computer program (Department of Biochemistry and Molecular Biology, University College, London, England).
24. Thornton, J. M., Edwards, M. S., Taylor, W. R. & Barlow, D. J. (1986) *EMBO J.* **5,** 409–413.
25. McDonald, I. K. & Thornton, J. M. (1994) *J. Mol. Biol.* **238,** 777–793.
26. Hurley, J. H., Thorness, P. E., Ramalingam, V., Helmers, N. H., Koshland, D. E. & Stroud, R. M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 8635–8639.
27. Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. & Hendrickson, W. A. (1991) *Science* **254,** 1608–1615.
28. Smith, D. L., Almo, S. C., Toney, M. D. & Ringe, D. (1989) *Biochemistry* **28,** 8161–8167.
29. Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M. & Wiley, D. (1990) *J. Mol. Biol.* **216,** 141–160.
30. Chothia, C. & Janin, J. (1975) *Nature (London)* **256,** 705–708.
31. Morgan, R. S., Miller, S. L. & McAdon, J. (1979) *J. Mol. Biol.* **127,** 31–39.
32. Connolly, M. L. (1986) *Biopolymers* **25,** 1229–1247.
33. Lawrence, M. C. & Colman, P. M. (1993) *J. Mol. Biol.* **234,** 946–950.
34. Vakser, I. A. & Aflalo, C. (1994) *Proteins: Struct. Funct. Genet.* **20,** 320–329.
35. Eigenbrot, C., Randal, M., Quan, C., Burnier, J., O'Connell, L., Rinderknecht, E. & Kossiakoff, A. A. (1991) *J. Mol. Biol.* **221,** 15–21.
36. Huber, R. (1979) *Trends Biochem. Sci.* **4,** 271–276.
37. Hubbard, S. J., Thornton, J. M. & Campbell, S. F. (1992) *Faraday Disc.* **43,** 13–23.
38. Navia, M. A., Fitzerald, P. M. D., Mckeever, B. M., C., L., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L. & Springer, J. P. (1989) *Nature (London)* **337,** 615–620.
39. Mitsui, Y., Satow, Y., Watanabe, Y., Hirono, S. & Iitaka, Y. (1979) *Nature (London)* **277,** 447–452.
40. Tainer, J. A., Getzoff, E. D., Beem, K. M., Richardson, J. S. & Richardson, D. C. (1982) *J. Mol. Biol.* **160,** 181–217.
41. Wilson, I. A. & Stanfield, R. L. (1993) *Curr. Opin. Struct. Biol.* **3,** 113–118.
42. Wilson, I. A. & Stanfield, R. L. (1994) *Curr. Opin. Struct. Biol.* **4,** 857–867.
43. Gerstein, M., Schulz, G. & Chothia, C. (1993) *J. Mol. Biol.* **229,** 494–501.
44. Gerstein, M., Anderson, B. F., Norris, G. E., Baker, E. N., Lesk, A. M. & Chothia, C. (1993) *J. Mol. Biol.* **234,** 357–372.
45. Davies, D. R. & Cohen, G. H. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7–12.
46. Stanfield, R. L., Takimoto-Kamimura, M., Rini, J. M., Profy, A. T. & Wilson, I. A. (1993) *Structure* **1,** 83–93.
47. Miller, S., Lesk, A. M., Janin, J. & Chothia, C. (1987) *Nature (London)* **328,** 834–836.