

# Macromolecular recognition in the Protein Data Bank

Joël Janin,<sup>a,b,\*</sup> Francis Rodier,<sup>a</sup>  
Pinak Chakrabarti<sup>c</sup> and Ranjit P.  
Bahadur<sup>b,c</sup>

<sup>a</sup>Laboratoire d'Enzymologie et de Biochimie Structurales, UPR9063, CNRS, 91198 Gif-sur-Yvette, France, <sup>b</sup>Institut de Biochimie et Biologie Moléculaire et Cellulaire, UMR8619, Bâtiment 430, Université Paris-Sud, 91405 Orsay, France, and <sup>c</sup>Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIII, Calcutta 700 054, India

Correspondence e-mail:  
joel.janin@ibbmc.u-psud.fr

Received 8 November 2005

Accepted 4 September 2006

Crystal structures deposited in the Protein Data Bank illustrate the diversity of biological macromolecular recognition: transient interactions in protein–protein and protein–DNA complexes and permanent assemblies in homodimeric proteins. The geometric and physical chemical properties of the macromolecular interfaces that may govern the stability and specificity of recognition are explored in complexes and homodimers compared with crystal-packing interactions. It is found that crystal-packing interfaces are usually much smaller; they bury fewer atoms and are less tightly packed than in specific assemblies. Standard-size interfaces burying 1200–2000 Å<sup>2</sup> of protein surface occur in protease–inhibitor and antigen–antibody complexes that assemble with little or no conformation changes. Short-lived electron-transfer complexes have small interfaces; the larger size of the interfaces observed in complexes involved in signal transduction and homodimers correlates with the presence of conformation changes, often implicated in biological function. Results of the CAPRI (critical assessment of predicted interactions) blind prediction experiment show that docking algorithms efficiently and accurately predict the mode of assembly of proteins that do not change conformation when they associate. They perform less well in the presence of large conformation changes and the experiment stimulates the development of novel procedures that can handle such changes.

## 1. Introduction

Protein–protein recognition is an essential component of the cell machinery implicated in most biological processes. Its study, a very active part of functional genomics, uses methods that range from genetics to biophysics, crystallography, bioinformatics and computer modelling. Here, we analyze structural data deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000), taking examples that illustrate both the specific recognition of proteins that form functional complexes and other multimolecular assemblies and the nonspecific interactions that build up protein crystals.

## 2. Diversity of protein–protein recognition

Protein–protein recognition is as diverse as life itself and the structures that are presently available cover only a fraction of that diversity (Jones & Thornton, 1995, 1996, 2000; Larsen *et al.*, 1998; Wodak & Janin, 2002; Noreen & Thornton, 2003a). The sample listed in Table 1 contains a few hundred hand-picked structures representing the various types of interactions that take place between proteins in cells and organisms. Protein–protein complexes comprise proteins that

**Table 1**

Sample used.

No., number of interfaces in the sample; Area  $B$ , interface area  $B = \text{ASA}_1 + \text{ASA}_2 - \text{ASA}_{12}$ ;  $f_{\text{nonpolar}}$ , fraction of the interface area contributed by nonpolar (carbon-containing) groups;  $f_{\text{buried}}$ , fraction of the interface atoms that have zero ASA in the complex;  $n_{\text{HB}}$ , number of hydrogen bonds;  $n_{\text{water}}/B$ , number of water molecules immobilized at the interface divided by the interface area (data from Rodier *et al.*, 2005).

Type of interface	No.	Area $B$ ( $\text{\AA}^2$ )	$f_{\text{nonpolar}}$	$f_{\text{buried}}$	$n_{\text{HB}}$	$n_{\text{water}}/B$ ( $10^{-3} \text{\AA}^{-2}$ )
Protein–protein complexes†						
Antigen–antibody	18	1660				
Protease–inhibitor	23	1900				
Signal transduction	11	2500				
Miscellaneous	18	1870				
All	70	1900	0.58	0.34	10.0	10.0
Electron-transfer complexes‡	10	1190	0.65			1.6
Homodimers§	122	3900	0.65	0.36	18.5	11.3
Crystal packing						
All interfaces¶	1320	570				
Large interfaces††	188	1510	0.58	0.21	5.4	15.3
Protein–DNA complexes‡‡	65	3180	0.52	0.27	21.8	

† Data from Chakrabarti & Janin (2002). ‡ Data from Table 2 of Crowley & Carrondo (2004) excluding the covalent complex adxadr. § Data from Bahadur *et al.* (2003). ¶ Data from Janin & Rodier (1995). †† Data from Bahadur *et al.* (2004). ‡‡ Data from Nadassy *et al.* (1999); the values of  $f_{\text{nonpolar}}$  and  $f_{\text{buried}}$  are for the protein side;  $f_{\text{nonpolar}} = 0.41$  on the DNA side.

fold independently; in contrast, homodimers are permanent assemblies, usually obligate as the isolated subunits are not present in cells. Crystal contacts are (very useful) artifacts of the crystallization procedure. Each complex or homodimer contains a specific interface, which we may compare with the nonspecific packing interfaces formed between protein molecules in a crystal. Crystal-packing interfaces are usually small, but we selected some that are comparable in size to specific interfaces. We analyze their geometric and chemical properties and attempt to draw general rules from the comparison. We confirm rules derived many years ago from analysis of only a handful of X-ray structures (Chothia & Janin, 1975; Janin & Chothia, 1990) and add some new ones (Lo Conte *et al.*, 1999), trusting that this trend will hold as more data shed further light on aspects of protein–protein recognition that still lack a structural basis.

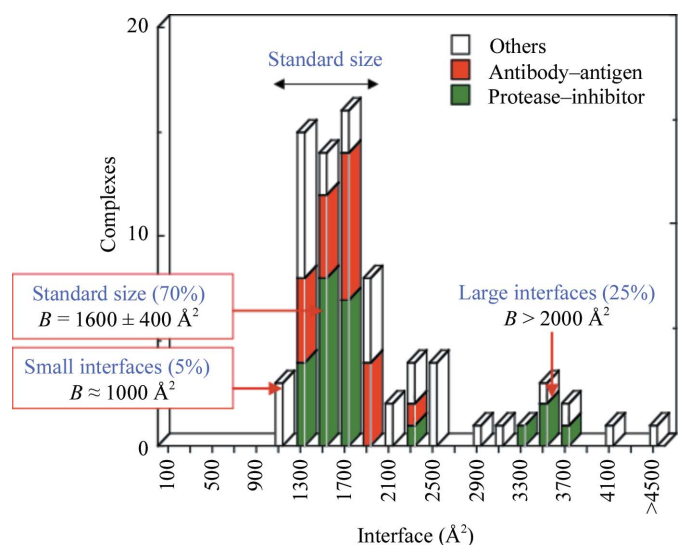
**2.1. Antigen–antibody and protease–inhibitor complexes: standard-size interfaces and rigid-body recognition**

The interaction of cognate antibodies with protein antigens plays an essential role in the immune system of vertebrates. It is a paradigm of specific recognition and one of the best represented in the PDB (Braden & Poljak, 2000; Sundberg & Mariuzza, 2002). Table 1 includes 18 protein antigen–antibody complexes and Fig. 1 shows the size distribution of their interfaces along with those of other transient complexes. The size is estimated as the interface area  $B = \text{ASA}_1 + \text{ASA}_2 - \text{ASA}_{12}$  calculated as the solvent-accessible surface area  $\text{ASA}_{12}$  of the complex less that of the dissociated components  $\text{ASA}_1$  and  $\text{ASA}_2$  (Lee & Richards, 1971; Chothia & Janin, 1975; note that other authors often report the quantity  $B/2$ ). The antibody and the antigen moieties of the complex contribute

almost equally to  $B$ . The distribution of the values is narrow for antigen–antibody complexes relative to the other types. All but one of the 19 interfaces are of ‘standard size’, with  $B$  in the range 1200–2000  $\text{\AA}^2$ . The average interface atom loses about 10  $\text{\AA}^2$  of ASA in the complex and therefore a standard-size interface involves about 80 atoms belonging to approximately 23 amino-acid residues on each component.

Another property shared by antigen–antibody complexes of known structure is the limited extent to which the antigen undergoes conformation changes upon binding the antibody. This can be assessed when the antigen structure is known independently, which is the case of most of the complexes in Table 1. A superposition of the free and bound antigen indicates that the main chain undergoes only local movements of 1–2  $\text{\AA}$ , although side-chain rotations occur. Thus, the antigen binds as a (quasi) rigid body, whereas some of the antigen-binding loops may be mobile in the antibody. However, the X-ray structure of the free antibody is rarely available and these changes can only be assessed in a few cases.

The size distribution of the protease–inhibitor interface is similar to that of the antigen–antibody complexes, but it is bimodal (Fig. 1). It has a major peak (19 complexes) centred at 1500  $\text{\AA}^2$  in which all interfaces are standard size (1200–2000  $\text{\AA}^2$ ) and a minor peak (four complexes) centred at 3500  $\text{\AA}^2$ . The complexes in the major peak include a variety of proteases and ‘canonical’ small inhibitors (Laskowski *et al.*, 2000); for example, the pancreatic trypsin inhibitor (PTI). In the minor peak, the inhibitors are larger molecules such as ornithodorin, which inhibits the serine protease thrombin. Ornithodorin is a duplicate of PTI and whilst its first PTI-like domain binds to the active site of thrombin in the same way as PTI does to trypsin, the second domain interacts at a different site (van de Loch *et al.*, 1996). Each interaction forms a recognition patch that buries approximately 1500  $\text{\AA}^2$  of



**Figure 1** The interface area of protein–protein complexes. Histogram of the values of the interface area  $B = \text{ASA}_1 + \text{ASA}_2 - \text{ASA}_{12}$  in 19 antigen–antibody complexes, 23 protease–inhibitor complexes and 33 other complexes. Interfaces with an area in the range 1200–2000  $\text{\AA}^2$  are labelled ‘standard size’. Adapted from Lo Conte *et al.* (1999).

protein surface. Thus, the interface comprises a single patch in trypsin–PTI, but two in thrombin–ornithodorin. All the protease complexes with small inhibitors and most if not all antigen–antibody complexes in our sample form single-patch interfaces (Chakrabarti & Janin, 2002). Moreover, a comparison of the complexes with the free proteases and free inhibitors indicates that rigid-body recognition is the rule in systems involving small inhibitors such as PTI, whereas the two-domain ornithodorin is likely to be highly flexible when free.

## 2.2. Electron-transfer complexes: a short-lived assembly with a small interface

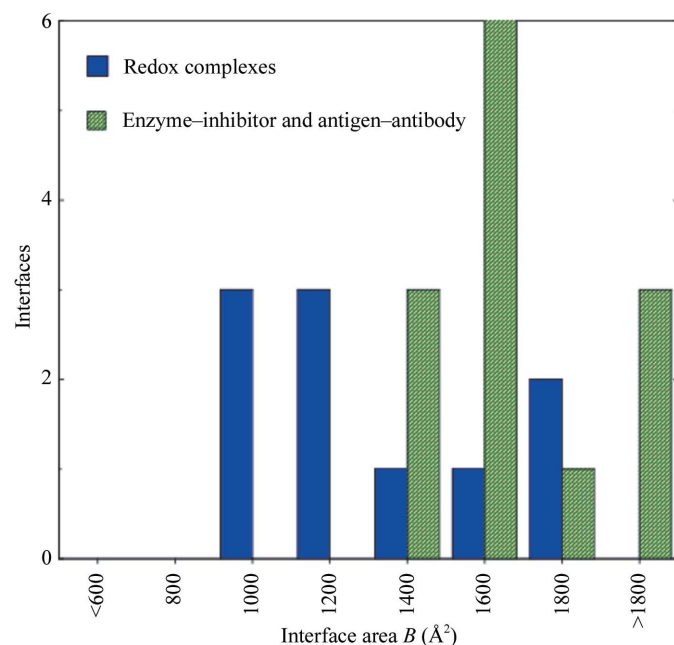
Once formed, the trypsin–PTI complex is extremely stable, with a dissociation constant  $K_d$  below picomolar and a half-life of months. Antigen–antibody complexes typically have a nanomolar  $K_d$  and half-lives of hours to days (Braden & Poljak, 2000; Sundberg & Mariuzza, 2002). These two categories are representative of the long-lived assemblies that have long been a favorite with biochemists and crystallographers. It is clear now that many biological processes rely on interactions that have a much shorter half-life and there is a shift in interest towards their study (Noreen & Thornton, 2003*b*). For instance, an enzyme that has a protein as substrate, a kinase for instance, must form a short-lived complex as the turnover rate  $k_{cat}$  is necessarily less than the rate of product dissociation.

Short-lived complexes are still few in the PDB and the best-represented category is the redox complex, where an electron

is transferred from one component to another. Electron transfer requires the donor and acceptor groups to be less than about 17 Å apart, but the geometry is not strict and transfer can take place at a high rate in loose complexes that dissociate in milliseconds and have a  $K_d$  above micromolar (Mathews *et al.*, 2000). Fig. 2, which uses data from Crowley & Carrondo (2004), indicates that their interfaces are often smaller than in protease–inhibitor or antigen–antibody complexes. Some electron-transfer complexes have standard-size interfaces, but a majority have  $B$  in the range 900–1200 Å<sup>2</sup>, which comprises only 5% of the sample of Fig. 1.

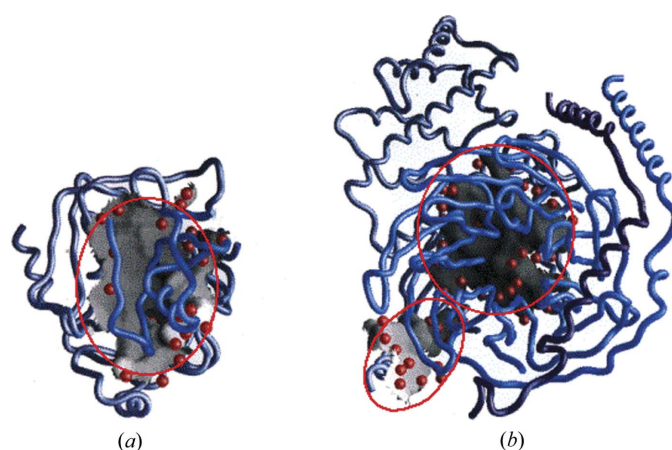
## 2.3. Signal transduction: flexible recognition and a large interface

The example of the short-lived redox complexes may suggest that stability is correlated with interface size, but if the correlation exists it does not extend to other categories. For instance, protein–protein complexes involved in signal transduction must assemble and dissociate in response to the cell environment. They cannot be long-lived, yet Table 1 shows that they often have large interfaces. An example is transducin, the trimeric G-protein coupled with rhodopsin in the retina. The visual signal is initiated when rhodopsin absorbs a photon and it is converted into a chemical signal (cyclic GMP) in a series of steps that include the dissociation of transducin into its  $G\alpha$  and  $G\beta\gamma$  components, GTP hydrolysis by  $G\alpha$  and the activation of guanylate cyclase by  $G\beta\gamma$ . The whole process is completed in milliseconds and therefore transducin dissociation must be fast.



**Figure 2**

Electron-transfer complexes have small interfaces. Values of the interface areas of redox complexes are taken from Table 2 of Crowley & Carrondo (2004), excluding the covalent complex adxadr. Those of the protease–inhibitor and antigen–antibody complexes are reported from Fig. 1 for comparison.



**Figure 3**

Flexible *versus* rigid-body recognition. (a) Chymotrypsin–ovomucoid: the protease surface in contact with the inhibitor (Icho; Fujinaga *et al.*, 1987) is viewed through the inhibitor main chain. This complex illustrates rigid-body recognition. It has a high affinity ( $K_d \approx 10^{-11}$  M), a standard-size interface ( $B = 1470$  Å<sup>2</sup>) in a single patch and a low  $C^\alpha$  r.m.s.d. between the free and bound components (0.6 Å). (b) Transducin: the surface of the  $G\alpha$  subunit that interacts with the  $G\beta\gamma$  pair of subunits is viewed through the  $G\beta\gamma$  main chain (PDB code 1got; Lambright *et al.*, 1996). Transducin illustrates flexible recognition; the  $G\alpha$ – $G\beta\gamma$  interface is large ( $B = 2500$  Å<sup>2</sup>), in two patches, and the association involves major conformation changes (1.8 Å  $C^\alpha$  r.m.s.d.) in both  $G\alpha$  and  $G\beta\gamma$ ; the affinity is comparatively low ( $K_d \approx 10^{-7}$  M).

Nevertheless, the  $G\alpha$ - $G\beta\gamma$  interface is larger ( $B = 2500 \text{ \AA}^2$ ) than in protease-inhibitor complexes that are much more stable. Fig. 3 compares this signal-transducing interface (PDB code 1got) with the standard-size interface of the chymotrypsin-ovomucoid complex (PDB code 1cho). In both pictures, one component (the protease or  $G\alpha$ ) is viewed through the backbone of the other component (ovomucoid or  $G\beta\gamma$ ) and the contact surface is imaged in grey. The chymotrypsin-ovomucoid interface forms a single patch approximately  $25 \times 20 \text{ \AA}$  in size, surrounded by a set of immobilized water molecules. In contrast, the  $G\alpha$ - $G\beta\gamma$  interface is in two patches, both lined with immobilized water. The major patch, similar in size to the chymotrypsin-ovomucoid interface, involves surface loops of both  $G\alpha$  and  $G\beta\gamma$  and the minor patch an  $\alpha$ -helix at the N-terminus of  $G\alpha$ .

The late Paul Sigler and his group at Yale University, who determined the X-ray structure of the transducin trimer (Lambright *et al.*, 1996), also determined those of  $G\alpha$  and  $G\beta\gamma$  independently. A comparison indicates that the surface loops of both components that interact to form the main patch move and change conformation in the complex. In addition, the helix of the minor patch is disordered in free  $G\alpha$ . Thus, large changes including a disorder-order transition accompany association in that system. Unlike trypsin-PTI and chymotrypsin-ovomucoid, which undergo rigid-body association,  $G\alpha$ - $G\beta\gamma$  recognition takes place between two flexible molecules. The free-energy cost of the conformation change is difficult to estimate, but we may assume that the presence of a large interface compensates for it.

The retina is a paradigm of signal transduction and transducin is a key element in it. The changes that occur within  $G\alpha$  and  $G\beta\gamma$  are an integral part of the process of signal transduction. By lowering the stability of the assembly, they enable the components of the trimeric G-protein to make other interactions; in  $G\alpha$ , they modulate the GTPase activity and nucleotide-binding affinity and they are also likely to contribute to the specific recognition of the activated form of rhodopsin.

#### 2.4. Protein-DNA recognition

Large conformation changes occur in a majority of the protein-protein complexes other than protease-inhibitor and antigen-antibody. In protein complexes with double-stranded DNA (Nadassy *et al.*, 1999; Jones *et al.*, 1999), conformation changes are the rule in the protein, the DNA or both. The average protein-DNA interface in a sample of 65 complexes has  $B$  near  $3200 \text{ \AA}^2$  (Table 1). The majority of the proteins in the sample are dimers and each subunit of the dimer forms an interface with  $B$  in the range  $1200$ - $2000 \text{ \AA}^2$ , the same as for standard-size protein-protein interfaces. Such an interface typically involves 24 amino-acid residues and 12 nucleotides. The smaller DNA-binding domains form standard-size interfaces and large proteins such as DNA polymerase form extensive interfaces with DNA. These proteins are multi-domain and different domains contribute separate recognition

patches to the interface (three in the case of DNA polymerase).

Flexibility in protein-DNA recognition is often part of the function. Examples are the bending of DNA by the lactose repressor, which enables the four subunits to bind at two different operator sites, or the large distortion in the double helix induced by the TATA box-binding protein, which guides the binding to DNA of other components of the transcription-initiation complex.

#### 2.5. Permanent assemblies versus transient complexes: interface size, composition and hydration

Unlike protein-protein and protein-DNA complexes, oligomeric proteins are usually permanent assemblies that only dissociate under denaturing conditions. Whereas most of the 122 homodimers selected by Bahadur *et al.* (2003) do not dissociate appreciably, Noreen & Thornton (2003*b*) assembled a set of weak homodimers that are in equilibrium with monomers. In either sample, all subunit interfaces have  $B > 900 \text{ \AA}^2$ . All but two of the weak interfaces reported in Noreen & Thornton (2003*b*) are standard size (noting that the reported value is  $\Delta\text{ASA} = B/2$ ). In the sample of Bahadur *et al.* (2003), only 20% of the interfaces are standard size; the remainder are larger, with  $B$  up to  $10\,000 \text{ \AA}^2$ .

The correlation between interface size and conformation changes seen in the complexes holds in homodimers, because oligomeric proteins are generally known to assemble while the subunits fold and folding is a disorder-to-order transition. In a complex, the protein surface involved in contacts remains solvent-accessible until the components meet and it cannot be very different from the remainder of the protein surface in its physical chemical properties. In a homodimer that assembles as it folds, the interface may be more like the protein interior. This shows in the values of  $f_{\text{nonpolar}}$  cited in Table 1;  $f_{\text{nonpolar}}$  is the fraction of the interface area  $B$  contributed by carbon-containing groups, which cannot hydrogen bond to water, as opposed to nitrogen- or oxygen-containing groups, which can. This fraction is 0.57 for the average solvent-accessible protein surface (Lee & Richards, 1971; Miller *et al.*, 1987), essentially the same as for the interfaces in complexes;  $f_{\text{nonpolar}}$  is higher (0.65) in homodimer interfaces and very low in protein-DNA interfaces (0.52 on the protein side, 0.41 on the DNA side). The high polarity of protein-DNA interfaces and the large number of hydrogen bonds in these complexes reflect the different chemistry of DNA and the large contribution of the phosphate backbone interacting with polar protein groups (Nadassy *et al.*, 1999).

The difference between transient and permanent assemblies is also apparent in the amino-acid composition of the interfaces, which resemble the protein surface in protein-protein complexes and the protein interior in homodimers (Janin *et al.*, 1988; Young *et al.*, 1994; Jones & Thornton, 1995; Tsai *et al.*, 1997; Bahadur *et al.*, 2003; Ofra & Rost, 2003). However, even in homodimers a majority of the atoms or residues that lose ASA remain solvent-accessible and therefore belong to the protein surface, not the interior. The values

of  $f_{\text{buried}}$  listed in Table 1 indicate that only about one-third of the atoms that contribute to  $B$  have zero ASA in the assembly. Defining the core of an interface as the set of residues that contain these buried atoms, Bahadur *et al.* (2003) find that the core is larger in homodimers than in complexes, but its amino-acid composition is the same. Interface cores are enriched in aromatic and aliphatic residues and depleted in the charged residues Lys, Glu and Asp, but not Arg. Moreover, interface core residues are better conserved in evolution than the rest of the protein surface (Guharoy & Chakrabarti, 2005) and they constitute the majority of the 'hot spots' which strongly destabilize the assembly when substituted by mutation (Covell & Wallqvist, 1997; Bogan & Thorn, 1998; Guerois *et al.*, 2002; Kortemme & Baker, 2002). Thus, interface cores resemble the protein interior except for the presence of Arg residues, which are as abundant at protein-protein interfaces, even in the core, as elsewhere on the protein surface.

Another major difference between protein-protein interfaces and the protein interior is the abundance of immobilized water (Larsen *et al.*, 1998; Janin, 1999; Rodier *et al.*, 2005). The interfaces of complexes and homodimers contain about ten immobilized water molecules per 1000 Å<sup>2</sup> of area (Table 1). As water is not consistently reported in medium-resolution X-ray structures and high-resolution structures contain at least 50% more, their actual number is probably in the range 10–20 per 1000 Å<sup>2</sup>. This amounts to 10–20% of the water molecules that are in contact with the protein surface before assembly (about one water per 10 Å<sup>2</sup> of ASA). Thus, dehydration of protein-protein interfaces is only 90–80% complete, whereas the protein interior is almost completely dehydrated (Hubbard & Argos, 1994). In transducin and the chymotrypsin-ovomucoid complex (Fig. 3), the interface core is dry and water is present only at the edge. This is the rule in homodimer interfaces, but immobilized water is often present throughout the interface in other systems (Rodier *et al.*, 2005).

### 3. Structural features associated with specificity

The noncovalent interactions that hold protein crystals together are the same as those that stabilize homodimers or complexes, but they are not subject to the biological selection that gives specificity to interactions involved in cellular functions. Thus, a comparison of the interfaces created by nonspecific contacts in crystals to the specific interfaces of homodimers and complexes should provide useful information on specificity.

#### 3.1. Crystal packing creates small interfaces with little conformation changes

From a structural point of view, there is an obvious difference between the two types of interfaces. The pairwise contacts in a protein crystal are much less extensive than in homodimers or complexes (Janin & Rodier, 1995; Dasgupta *et al.*, 1997; Janin, 1997; Carugo & Argos, 1997). The average area of crystal-packing interfaces is only 570 Å<sup>2</sup> per interface (Fig. 4), yet the total buried surface is large because each

protein molecule makes 6–12 such interfaces (Janin & Rodier, 1995). Moreover, with few exceptions, proteins maintain the same structure to within 1 Å r.m.s.d. for the main chain in different crystal forms, suggesting that crystal packing has little effect on conformation.

Nevertheless, some crystals contain pairwise interfaces comparable in size to those of protein-protein complexes. The tail of the distribution in Fig. 4 comprises large packing interfaces, often associated with twofold symmetry elements forming 'crystal dimers' that may be mistaken for real dimers. To identify structural features other than size that distinguish between the two, we selected from a set of crystals of monomeric proteins 188 packing interfaces with  $B > 800$  Å<sup>2</sup> including 105 with twofold symmetry (Bahadur *et al.*, 2004). Table 1 shows that on average these large crystal-packing interfaces are standard size and have the same  $f_{\text{nonpolar}}$  fraction as in complexes. As homodimers have a larger average  $f_{\text{nonpolar}}$ , the chemical composition of the interface may distinguish between real and crystal dimers, but the distributions of individual values overlap. In the same way, the amino-acid compositions differ, but not so much as to remove ambiguities between the two kinds of interfaces (Ponstingl *et al.*, 2000; Ofran & Rost, 2003; Bahadur *et al.*, 2004).

#### 3.2. The atomic packing of specific versus non-specific interfaces

On the other hand, the fraction  $f_{\text{buried}}$  of the interface atoms that have zero ASA in the assembly is much lower at crystal contacts than in homodimers or complexes (Table 1). As most of their atoms remain in contact with water, crystal contacts contain relatively fewer hydrogen bonds (about one per 280 Å<sup>2</sup> instead of one per 200 Å<sup>2</sup>). The reason for the lack of buried atoms is obvious in Fig. 5, which represents the interfaces of a real homodimer (PDB code 1kba) and a crystal dimer (PDB code 1qci). This particular homodimer has a small interface with about the same number of atoms as in the crystal-packing interface, but these atoms group together, whereas they spread apart in the crystal dimer.

The packing of atoms at an interface may be quantified by measuring the Voronoi atomic volumes and comparing their values with those of the protein interior (Richards, 1974; Harpaz *et al.*, 1994; Gerstein *et al.*, 1995). This measurement indicates that the packing at the interfaces of protein-protein complexes is as tight as inside proteins (Lo Conte *et al.*, 1999). However, the Voronoi construction requires the atoms to be buried, which only 21% are at crystal-packing interfaces. Other parameters may be used to compare these interfaces with those of complexes and homodimers:  $L_d$  and  $G_d$ , which measure the number density of atoms at an interface (Bahadur *et al.*, 2004),  $I_{\text{gap}}$ , the volume of the gap left between the subunits divided by the interface area (Laskowski, 1995), and  $S_c$ , an index of the shape complementarity of their surfaces (Lawrence & Colman, 1993). In Fig. 6, these four parameters are normalized to their mean value in homodimers and the normalized values are all smaller for crystal contacts than specific interfaces.



3.3. In a protein crystal, which contacts are biologically meaningful and which are not?

The PDB is full of entries describing proteins of unknown or uncertain quaternary structure. Although quaternary structure should be determined in solution, criteria based on atomic coordinates can be useful, especially when the protein is known to be oligomeric and the crystal packing offers more than one possible mode of assembly. In such cases, the specific interfaces that build the oligomer must be identified among other interfaces that do not exist in solution.

The Probable Quaternary Structure database (PQS; Henrick & Thornton, 1998) assigns quaternary structures based on the interface size and amino-acid composition (Ponstingl *et al.*, 2000). The accuracy of dimer prediction, only about 80%, is much improved when multiple sequences are available and conservation can be estimated (Valdar & Thornton, 2001). Still, we find that geometry alone can do better than 80%. In the sample of Table 1, two protein molecules related by twofold symmetry and burying more than  $B = 800 \text{ \AA}^2$  have an 88% probability of being a homodimer in solution if

$$(f_{\text{nonpolar}} B > 1000 \text{ \AA}^2 \text{ and } f_{\text{buried}} > 0.24)$$

$$\text{or } (f_{\text{nonpolar}} > 0.61 \text{ and } n_{\text{HB}} > 8)$$

and they have the same 88% probability of being a crystal dimer (an artefact of crystallization) if these criteria are not met (this application of the *WizWhy* data-mining algorithm is courtesy of Dr M. Marcus-Kalish, Tel Aviv University).

4. Predicting interaction: the CAPRI experiment

Structural genomics programs perform high-throughput structure determination of individual gene products. They aim to produce a set of structures that can be extended to all proteins by homology modelling. Whereas multi-protein complexes can also be approached in a systematic manner, they are difficult to prepare and crystallize and their study is not prone to high throughput. Thus, there is a great interest in predicting the mode of association of proteins based on their structure. The CAPRI (critical assessment of predicted interactions) experiment was designed to evaluate the capacity of docking algorithms to perform such predictions (Janin *et al.*, 2003).

A CAPRI target is a complex between proteins of known structure and for which a X-ray structure has been determined, but not yet published. A round of prediction opens each time a target is offered by crystallographers to the CAPRI Management Committee. Predictors are given PDB entry codes for the proteins and they have a few

weeks to submit up to ten models of the complex to the CAPRI web site (<http://capri.ebi.ac.uk>; managed by Dr K. Henrick, Hinxton, England). These models are then assessed by comparison to the X-ray structure and ranked based on two types of criteria: the identification and location of the binding epitopes and the capacity of the models to reproduce the residue-residue contacts seen at the interface of the experi-

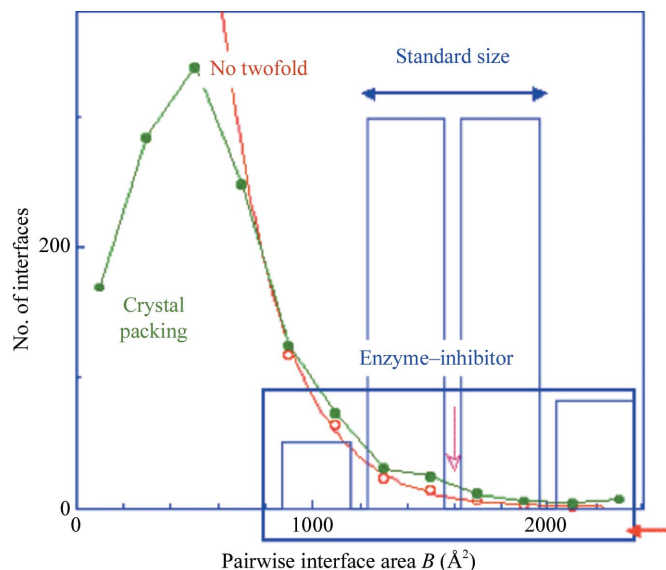


Figure 4 The size of crystal-packing interfaces. Distribution of the interface area  $B$  of 1320 pairs of molecules in crystals of monomeric proteins analyzed by Janin & Rodier (1995). The average is  $B = 570 \text{ \AA}^2$ . Interfaces with values of  $B$  comparable to those found in protein-protein complexes ( $B > 800 \text{ \AA}^2$ ) occur mostly in crystals with twofold symmetry; they form ‘crystal dimers’ that may be mistaken for real homodimers. In crystals with no twofold symmetry, the number of packing interfaces with  $B > 800 \text{ \AA}^2$  follows an extreme value distribution, approximated here by the red line (Janin, 1997). The boxed region includes 103 crystal dimers and 85 other large crystal-packing interfaces whose properties may be compared with those of the interfaces in complexes and homodimers (Bahadur *et al.*, 2004).

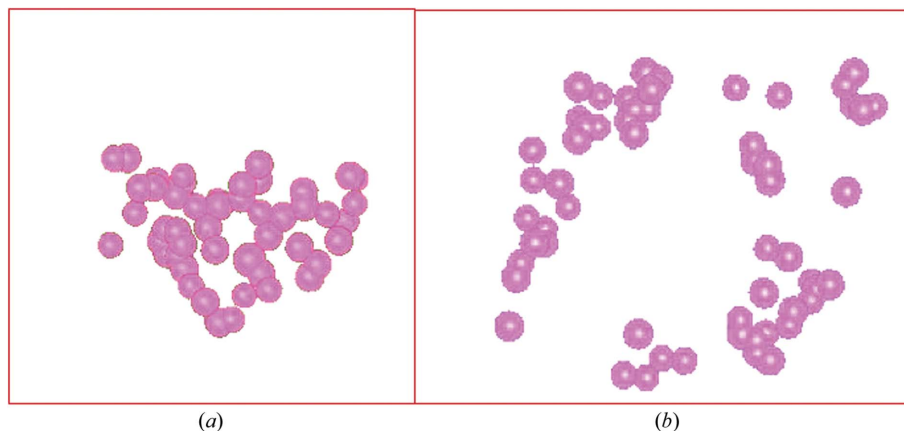
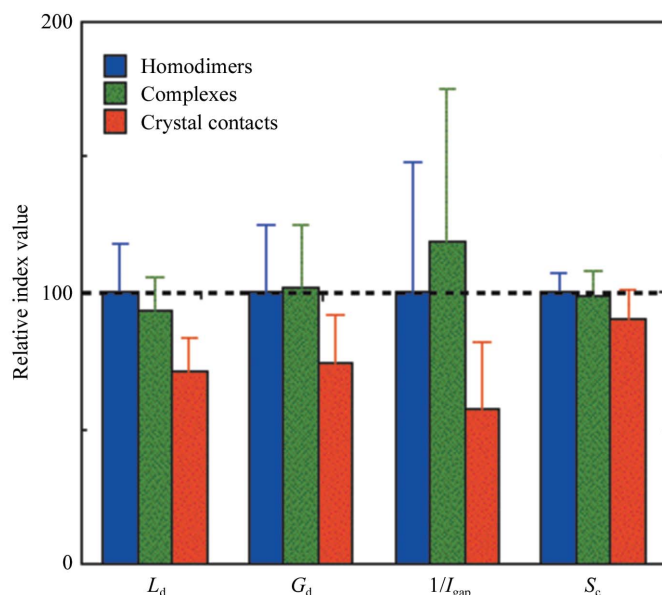


Figure 5 Atomic packing at a specific and a nonspecific interface. The pink balls represent atoms at the interface of (a) one subunit of the  $\kappa$ -bungarotoxin dimer (PDB code 1kba) and (b) the pokeweed protein crystal dimer (PDB code 1qci). Each interface buries approximately  $1000 \text{ \AA}^2$  and includes about 50 atoms, but whereas these atoms form a compact patch in the case of the homodimer, they spread apart in the crystal contact.



**Figure 6**

Interface compactness from four packing indices. The packing indices  $L_d$ ,  $G_d$ ,  $1/I_{gap}$  and  $S_c$  are calculated for the interfaces of Table 1.  $L_d$  and  $G_d$  are the local and global atomic density indices of Bahadur *et al.* (2004);  $I_{gap}$  is the gap-volume index of Laskowski (1995);  $S_c$  is the shape-complementarity score of Lawrence & Colman (1993). The packing indices are normalized relative to their average value in homodimer interfaces. Crystal contacts yield values of all four indices that are less than 100, indicating that their packing is less tight than in the specific interfaces.

mental structure. In 2001–2005, 21 target complexes have been subjected to prediction by up to 37 predictor groups. They include antigen–antibody and enzyme–inhibitor complexes and also complexes involved in a variety of cellular processes including signal transduction and oligomeric proteins (Janin, 2005).

The results of the predictions have been evaluated by the CAPRI assessors (Professor S. Wodak and her colleagues in Brussels and Toronto) and published in detail for Rounds 1–5 (Mendez *et al.*, 2003, 2005). 14 of the 19 targets had predictions that were judged ‘good’ or ‘high-quality’ by the assessors. The binding epitopes were within 2 Å of their correct position and the models reproduced at least 30% of the native residue–residue contacts. Such models can easily be tested by biochemical methods and guide site-directed mutagenesis in the absence of an experimental structure. In some cases, many predictor groups who used different docking algorithms and different scoring functions submitted models in which most of the interface residues were correctly identified and the geometry approximately correct.

Prediction was of lesser quality on two targets and it failed entirely on three. The one failure in the period 2004–2005 was for a homodimeric protein in which the subunits change their mode of association upon a mutation. It undergoes large domain movements that proved to be beyond the reach of existing algorithms, although smaller changes were successfully simulated in other targets. In general, the quality of the predictions was highly dependent on the extent to which the target differed in structure from the free proteins. When the

conformation changes were small, some very accurate predictions were made. Moreover, a comparison with the results of the first CAPRI rounds indicates that obvious progress has been made, especially in handling conformation changes and in ranking docking solutions (Mendez *et al.*, 2005).

## 5. Conclusion

The systems we have analyzed are diverse in structure and function, yet they have features in common. One is the size of the interface: to make a biologically relevant protein–protein (or protein–DNA) interaction, about 900 Å<sup>2</sup> of macromolecular surface contributed by some 45 atoms of each component must be removed from contact with the solvent. It was suggested many years ago that this lower limit reflects the physics of macromolecular association (Chothia & Janin, 1975). When two molecules associate, translational/rotational degrees of freedom are converted into vibrational degrees of freedom of lower entropy. The free-enthalpy cost of the conversion  $\Delta G_{rt}$  has been disputed and estimates range from 0 to 60 kJ mol<sup>-1</sup> in the 1 M standard state.  $\Delta G_{rt}$  may be small in redox complexes that have a loose geometry, but the higher value is likely to apply to antigen–antibody or protease–inhibitor complexes and to homodimers (Finkelstein & Janin, 1989; Karplus & Janin, 1999). Thus, a micromolar affinity that implies  $\Delta G^\circ = -RT \ln K_d \simeq 34$  kJ mol<sup>-1</sup> may require 94 kJ mol<sup>-1</sup> of favourable free enthalpy. Most of that free enthalpy arises at the interface: from the hydrophobic effect as nonpolar groups become dehydrated, close-packed atoms that optimize van der Waals forces and interactions between polar protein groups that replace weaker interactions with water (Nicholls *et al.*, 1991; Young *et al.*, 1994; Xu *et al.*, 1997; Janin, 2000). On average, each of these terms scales with the interface area. Thus, interfaces with  $B \simeq 1000$  Å<sup>2</sup> may allow short-lived assemblies only, whereas those with  $B$  in the 1200–2000 Å<sup>2</sup> range provide the stability and the high specificity exemplified by antigen–antibody recognition.

Complexes that form standard-size interfaces generally assemble by a rigid-body recognition mechanism. The macromolecular surfaces involved are complementary in shape and physical chemical properties and need only local adjustments to optimize their interaction. In contrast, transducing and protein–DNA complexes undergo flexible recognition in which the two surfaces must rearrange to achieve complementarity. The two mechanisms relate to different biological functions and the large conformation changes that accompany binding often play a critical role in function.

Docking algorithms reproduce rigid-body recognition efficiently and accurately, but they perform less well on systems that undergo large conformation changes. Simulating flexible recognition is a very active field of study and the CAPRI experiment has been a powerful incentive for developing novel procedures leading to significant progress in the last 4 y. Predictor groups wish very much that the experiment will continue. This will require new targets to be offered by crystallographers, who we thank in advance for their coop-

eration: structural biologists, please provide targets for CAPRI!

This work has benefitted over many years from the collaboration of Dr Cyrus Chothia (Cambridge) and Shoshana Wodak (Free University of Brussels and University of Toronto). JJ acknowledges financial support from the EIDIPP program of Action Concertée Incitative IMPBio.

## References

- Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2003). *Proteins*, **53**, 708–719.
- Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). *J. Mol. Biol.* **336**, 943–955.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
- Bogan, A. A. & Thorn, K. S. (1998). *J. Mol. Biol.* **280**, 1–9.
- Braden, B. C. & Poljak, R. J. (2000). *Protein–Protein Recognition*, edited by C. Kleanthous, pp. 126–161. Oxford University Press.
- Carugo, O. & Argos, P. (1997). *Protein Sci.* **6**, 2261–2263.
- Chakrabarti, P. & Janin, J. (2002). *Proteins*, **47**, 334–343.
- Chothia, C. & Janin, J. (1975). *Nature (London)*, **256**, 705–708.
- Covell, D. G. & Wallqvist, A. (1997). *J. Mol. Biol.* **269**, 281–297.
- Crowley, P. B. & Carrondo, M. A. (2004). *Proteins*, **55**, 603–612.
- Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. & Bell, J. A. (1997). *Proteins*, **28**, 494–514.
- Finkelstein, A. V. & Janin, J. (1989). *Protein Eng.* **3**, 1–3.
- Fujinaga, M., Sielecki, R., Read, R. J., Ardelt, W., Laskowski, M. & James, M. N. J. (1987). *J. Mol. Biol.* **195**, 397–418.
- Gerstein, M., Tsai, J. & Levitt, M. (1995). *J. Mol. Biol.* **249**, 955–966.
- Guerois, R., Nielsen, J. E. & Serrano, L. (2002). *J. Mol. Biol.* **320**, 369–387.
- Guharoy, M. & Chakrabarti, P. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.
- Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.
- Henrick, K. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 358–361.
- Hubbard, S. J. & Argos, P. (1994). *Protein Sci.* **3**, 2194–2206.
- Janin, J. (1997). *Nature Struct. Biol.* **4**, 973–974.
- Janin, J. (1999). *Structure*, **7**, R277–R279.
- Janin, J. (2000). *Protein–Protein Recognition*, edited by C. Kleanthous, pp. 1–32. Oxford University Press.
- Janin, J. (2005). *Proteins*, **60**, 170–175.
- Janin, J. & Chothia, C. (1990). *J. Biol. Chem.* **265**, 16027–16030.
- Janin, J., Henrick, K., Moulton, J., Ten Eyck, L., Sternberg, M., Vajda, S., Wakser, I. & Wodak, S. J. (2003). *Proteins*, **52**, 2–9.
- Janin, J., Miller, S. & Chothia, C. (1988). *J. Mol. Biol.* **204**, 155–164.
- Janin, J. & Rodier, F. (1995). *Proteins*, **23**, 580–587.
- Jones, S. & Thornton, J. M. (1995). *Prog. Biophys. Mol. Biol.* **63**, 31–65.
- Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
- Jones, S. & Thornton, J. M. (2000). *Protein–Protein Recognition*, edited by C. Kleanthous, pp. 33–59. Oxford University Press.
- Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). *J. Mol. Biol.* **287**, 877–896.
- Karplus, M. & Janin, J. (1999). *Protein Eng.* **12**, 185–186.
- Kortemme, T. & Baker, D. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
- Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E. & Sigler, P. B. (1996). *Nature (London)*, **379**, 311–319.
- Larsen, T. A., Olson, A. J. & Goodsell, D. S. (1998). *Structure*, **6**, 421–427.
- Laskowski, M. Jr, Qasim, M. A. & Lu, S. M. (2000). *Protein–Protein Recognition*, edited by C. Kleanthous, pp. 228–279. Oxford University Press.
- Laskowski, R. A. (1995). *J. Mol. Graph.* **13**, 323–330.
- Lawrence, M. C. & Colman, P. M. (1993). *J. Mol. Biol.* **234**, 946–950.
- Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.
- Locht, A. van de, Stubbs, M. T., Bode, W., Friedrich, T., Bollschweiler, C., Hoffken, W. & Huber, R. (1996). *EMBO J.* **15**, 6011–6017.
- Lo Conte, L., Chothia, C. & Janin, J. (1999). *J. Mol. Biol.* **285**, 2177–2198.
- Mathews, F. S., Mauk, A. G. & Moore, G. R. (2000). *Protein–Protein Recognition*, edited by C. Kleanthous, pp. 60–101. Oxford University Press.
- Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. (2003). *Proteins*, **52**, 51–67.
- Mendez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. (2005). *Proteins*, **60**, 150–169.
- Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). *J. Mol. Biol.* **196**, 641–656.
- Nadassy, K., Wodak, S. J. & Janin, J. (1999). *Biochemistry*, **38**, 1999–2017.
- Nicholls, A., Sharp, K. & Honig, B. (1991). *Proteins*, **11**, 281–296.
- Noreen, I. M. & Thornton, J. M. (2003a). *EMBO J.* **22**, 3486–3492.
- Noreen, I. M. & Thornton, J. M. (2003b). *J. Mol. Biol.* **325**, 991–1018.
- Ofran, Y. & Rost, B. (2003). *J. Mol. Biol.* **325**, 377–387.
- Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). *Proteins*, **41**, 47–57.
- Richards, F. M. (1974). *J. Mol. Biol.* **82**, 1–14.
- Rodier, F., Bahadur, R. P., Chakrabarti, P. & Janin, J. (2005). *Proteins*, **60**, 30–45.
- Sundberg, E. J. & Mariuzza, R. A. (2002). *Adv. Protein Chem.* **61**, 119–160.
- Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). *Protein Sci.* **6**, 53–64.
- Valdar, W. S. & Thornton, J. M. (2001). *J. Mol. Biol.* **313**, 399–416.
- Wodak, S. J. & Janin, J. (2002). *Adv. Protein Chem.* **61**, 9–73.
- Xu, D., Tsai, C. J. & Nussinov, R. (1997). *Protein Eng.* **10**, 999–1012.
- Young, L., Jernigan, R. L. & Covell, D. G. (1994). *Protein Sci.* **3**, 717–729.