Quarterly Reviews of Biophysics **41**, 2 (2008), pp. 133–180. © 2008 Cambridge University Press
doi:10.1017/S0033583508004708 Printed in the United Kingdom

Protein–protein interaction and quaternary structure

Joël Janin[1]*, Ranjit P. Bahadur[2] and Pinak Chakrabarti[3]

[1] Yeast Structural Genomics, IBBMC UMR 8619 CNRS, Université Paris-Sud, Orsay, France
[2] School of Engineering and Science, Jacobs University Bremen, Bremen, Germany
[3] Department of Biochemistry, Bose Institute, Calcutta, India

Abstract. Protein–protein recognition plays an essential role in structure and function. Specific non-covalent interactions stabilize the structure of macromolecular assemblies, exemplified in this review by oligomeric proteins and the capsids of icosahedral viruses. They also allow proteins to form complexes that have a very wide range of stability and lifetimes and are involved in all cellular processes. We present some of the structure-based computational methods that have been developed to characterize the quaternary structure of oligomeric proteins and other molecular assemblies and analyze the properties of the interfaces between the subunits. We compare the size, the chemical and amino acid compositions and the atomic packing of the subunit interfaces of protein–protein complexes, oligomeric proteins, viral capsids and protein–nucleic acid complexes. These biologically significant interfaces are generally close-packed, whereas the non-specific interfaces between molecules in protein crystals are loosely packed, an observation that gives a structural basis to specific recognition. A distinction is made within each interface between a core that contains buried atoms and a solvent accessible rim. The core and the rim differ in their amino acid composition and their conservation in evolution, and the distinction helps correlating the structural data with the results of site-directed mutagenesis and *in vitro* studies of self-assembly.

* Author for correspondence: J. Janin, Yeast Structural Genomics, IBBMC UMR 8619 Université Paris-Sud, 91405 Orsay, France.
Tel.: +33 1 69 15 79 66; Fax: +33 1 69 85 37 15; Email: joel.janin@u-psud.fr

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

1. Introduction

Most proteins are made of more than one polypeptide chain, and thus they have a quaternary structure (QS) in the classical nomenclature of Linderström-Lang & Schellman (1959), who named primary structure the amino acid sequence, secondary structure, the α helices and β sheets, and tertiary structure, the chain fold. Moreover, many, if not all, proteins interact with others to form binary complexes or higher-order assemblies that carry out all types of cellular processes. Indeed, the biological function of a protein can be seen as defined by the context of its interactions in the cell, and inappropriate interactions can lead to diseases (Alberts, 1998; Eisenberg *et al.* 2000). Thus, the unraveling of the underlying principles that govern protein–protein recognition is both central to the construction of the networks that define cell biology (Robinson *et al.* 2007) and instrumental in new drug development (Wells & McClendon, 2007).

The quaternary structure is a very early discovery in comparison with other levels of macromolecular assembly in biology. It was first identified in the mid 1920s by Svedberg (1927), when he determined the molecular weight of hemoglobin by sedimentation in the ultracentrifuge. The value he obtained, almost 68 000 Da, implied the presence of four subunits in the molecule. Sedimentation also showed that hemocyanin, a copper-containing protein, had a molecular

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

weight of millions, and presumably many subunits. Svedberg's discovery predates by decades that of the α helix and the β sheet by Pauling & Corey (1951), the first amino acid sequence of Sanger & Thompson (1953) and the first X-ray structure of Perutz (1960). Perutz' crystallographic studies of hemoglobin revealed the secondary and tertiary structures of the subunits, fully confirmed Svedberg's description of its QS and showed that the QS changes when oxygen binds. They inspired Monod and his collaborators, who introduced the concept of allostery. The allosteric model of regulatory mechanisms gives a central role to the QS and the way it changes when ligands bind (Monod *et al.* 1963, 1965). In those years, only a few scores of proteins had their sequence or X-ray structure determined, but many had their QS established, mostly in the ultracentrifuge, so that Darnall & Klotz (1975) could tabulate the QS of more than 500 proteins. The advent of DNA sequencing changed the setting of protein studies altogether. Obtaining an amino acid sequence suddenly became easy and fast, and a wide gap opened between our knowledge of the primary structure and that of the other levels of protein structure. Structural genomics initiatives, launched worldwide in 1998–2000 to close that gap, initially targeted single-gene products (Sali, 1998), a choice that reflects the views of that time. Since then, genome-wide genetic and biochemical studies have demonstrated that most gene products are part of multi-molecular assemblies in all cells and organisms (Giot *et al.* 2003; Li *et al.* 2004; Gavin *et al.* 2006; Krogan *et al.* 2006). Protein–protein interaction and QS have now returned to the front of the stage, and protein assemblies are the targets of several recent structural genomics initiatives (Russell *et al.* 2004; Janin, 2007), and structural biologists make major efforts to study them by crystallography, nuclear magnetic resonance (NMR) and electron microscopy.

The QS of a protein or a protein assembly is almost invariably essential to its function, and it must be established along with the sequence and fold of its components. This implies determining first the subunit composition, then the geometry of the assembly, and especially its symmetry, and lastly, the details of the interactions made by chemical groups and amino acid residues at the interfaces between the subunits. This review is devoted to the analysis of such interactions in different types of assemblies for which high-resolution structural data are available from X-ray studies. Protein–protein complexes are non-obligate, and mostly transient, assemblies that form when two preformed proteins meet. Oligomeric proteins assemble as the constituent polypeptide chains fold, and are mostly permanent; as their name (coined by Monod) implies, they comprise a few subunits. Icosahedral virus capsids are also permanent, but they comprise tens to hundreds of subunits. Whereas X-ray studies usually leave the nucleic acid component of icosahedral viruses undefined, a comparison of protein–protein interaction with protein–DNA and protein–RNA interaction is of general interest, and we include here data on all three processes of macromolecular recognition.

Since Svedberg, hemoglobin has been the paradigm oligomeric protein. Mammalian hemoglobins are heterotetramers, 'hetero' referring to the different amino acid sequences of the α and β chains. Their QS can be noted as $\alpha_2\beta_2$ or $(\alpha\beta)_2$ to show that they comprise two $\alpha\beta$ pairs related by twofold symmetry. The pairs are oriented differently in deoxy and in oxy-hemoglobin, which affects their interface and leads to the change in heme affinity for oxygen that makes oxygen binding cooperative (Perutz, 1970; Baldwin & Chothia, 1979). Most animal species have hemoglobins. Their sequences are related and they have the same characteristic fold, but not necessarily the same QS: some are homodimers (the two chains have the same sequence), others are monomers, or form larger assemblies. Their function is to bind oxygen in all cases, but the diversity of the QS allows oxygen binding to be regulated in different manners adapted to the

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Table 1. *Databases and Web servers for structure-based protein–protein interactions*

3D Complex	http://3dcomplex.org/
3DID	http://gatealoy.pcb.ub.es/3did/
ASEdb	http://nic.ucsf.edu/asedb
CAPRI	http://capri.ebi.ac.uk/
ClusPro	http://nrc.bu.edu/cluster/
ConSurf	http://consurf.tau.ac.il
Dockground	http://dockground.bioinformatics.ku.edu/
ExPASy	http://www.expasy.ch/
GRAMM-X	http://vakser.bioinformatics.ku.edu/resources/gramm/grammx
HADDOCK	http://haddock.chem.uu.nl/
InterPreTS	http://www.russell.embl.de/cgi-bin/interprets2
Interpro	http://www.ebi.ac.uk/interpro/
Intervor	http://bombyx.inria.fr/Intervor/intervor.html
Ipfam	http://www.sanger.ac.uk/Software/Pfam/iPfam/
MultiDock	http://www.sbg.bio.ic.ac.uk/docking/multidock.html
PatchDock	http://bioinfo3d.cs.tau.ac.il/PatchDock
PDB	http://www.rcsb.org/pdb/
PFAM	http://www.sanger.ac.uk/Software/Pfam/
PIbase	http://alto.compbio.ucsf.edu/pibase/
PiQSi	http://www.piqsi.org/
PISA	http://www.ebi.ac.uk/msd-srv/prot_int/
PITA	http://www.ebi.ac.uk/thornton-srv/databases/pita/
PP	http://www.biochem.ucl.ac.uk/bsm/PP/server/
PQS	http://pqs.ebi.ac.uk/
PRISM	http://prism.ccbb.ku.edu.tr/prism/
ProFace	http://www.boseinst.ernet.in/resources/bioinfo/stag.html
ProtBuD	http://dunbrack.fccc.edu/ProtBuD/
RosettaDock	http://rosettadock.graylab.jhu.edu/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
Scorecons	http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl
SKE-Dock	http://www.pharm.kitasato-u.ac.jp/biomoleculardesign/files/ske_dock.htm
SmoothDock	http://structure.pitt.edu/servers/smoothdock/
SymmDock	http://bioinfo3d.cs.tau.ac.il/SymmDock/
VIPERdb	http://viperdb.scripps.edu/

physiology of each organism. A number of oligomeric proteins and protein–protein complexes have regulatory properties like hemoglobin. In these proteins and many others, the function directly implicates the subunit interactions. Thus, the antigen binding site of an immunoglobulin is shared between the heavy and light chains, the active site of an oligomeric enzyme can be at a subunit interface, and molecular machines work by changing subunit–subunit contacts in a cyclic manner; ATP synthase (Stock *et al.* 1999, 2000) is a well-known example. An understanding of their biological function depends on analyzing their structure and the interactions between their subunits.

The renewed interest in protein–protein interaction has led to the publication in recent years of several major reviews (Noreen & Thornton, 2003a; Ponstingl *et al.* 2005; Janin *et al.* 2007) and collective books (Kleanthous, 2000; Janin & Wodak, 2003; Fu, 2004). New tools have been developed for its study in domains that range from genetics, cell biology and biochemistry to analytical chemistry, biophysics and structural biology. The emphasis in this review is on structure-based bioinformatics and computational tools, and especially the tools that are publicly available as Web servers (URLs are cited in Table 1). We review here results of their application to sets of complexes, oligomeric proteins and viral capsids, which illustrate the role of protein–protein

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

interaction in a wide variety of biological processes. We also introduce for comparison data on two systems of a different nature: protein–nucleic acid complexes and protein crystals. The first illustrates how the chemical nature of the partners modulates macromolecular interactions, and the second sheds light on the structural basis of specificity, an essential feature of biological assemblies that crystal packing lacks. The general rules that can be drawn from that analysis are relevant to the nature, the stability and the specificity of the interactions and shed light on protein evolution and the manner in which biological macromolecules self-assemble.

2. Tools to study quaternary structure

2.1 Experimental determination of the subunit composition

To determine the QS of a protein, one needs first to know its subunit composition. This may be established by introducing chemical cross-links between the polypeptide chains, or, more commonly, by comparing the molecular weights of the native protein and the constituent chains. The subunit molecular weights are obtained by gel electrophoresis under denaturing conditions, or calculated from the amino acid sequence taking into account post-translational modifications, if any. They can also be accurately measured by mass spectrometry, a powerful method not easily applicable to the native protein at present, because non-covalent bonds tend to break during sample desorption. Appropriate procedures are actively developed, and mass spectrometry will certainly be in a near future the choice method to determine the QS of proteins (Benesch & Robinson, 2006).

At present, native molecular weights are commonly measured directly by equilibrium analytical centrifugation, static light scattering or small-angle X-ray scattering (SAXS), methods that require purified protein in milligram quantity, or indirectly by methods less demanding in terms of equipment and the protein sample. Dynamic light scattering (DLS) measurements of the translational diffusion coefficient, and NMR or fluorescence polarization measurements of the rotational diffusion coefficient, yield data from which a molecular weight can be derived if the protein is known to be globular. Gel filtration on a molecular sieve (also called size exclusion chromatography), the most common method of all, yields the Stokes radius of the protein. Since the diffusion coefficients and the Stokes radius depend on the shape of the protein as well as its size, a QS based on a gel filtration pattern or a DLS measurement may not be correct and it should be verified by other methods.

Most non-obligate complexes, and a few oligomeric proteins, dissociate at low concentration. This can be seen in the ultracentrifuge, or by gel filtration or DLS, when the dissociation constant of the same order as the protein concentration, or typically 10^{-6}–10^{-4} M in such studies. However, a heterogeneous sample may yield a similar pattern to a monomer–oligomer equilibrium, and the measurement has to be made at several concentrations in order to demonstrate that the system is at thermodynamic equilibrium. With non-obligate complexes, the equilibrium can also be analyzed after mixing the components, but this is not generally feasible with oligomeric proteins, very few of which are available in a monomeric form.

2.2 Molecular symmetry of oligomeric proteins

A protein with *n* identical subunits usually has internal symmetry. The symmetry operations that superimpose an object onto itself form a point group. Mirror symmetries being excluded for proteins, the point group can be of one of three types: cyclic, dihedral or cubic

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

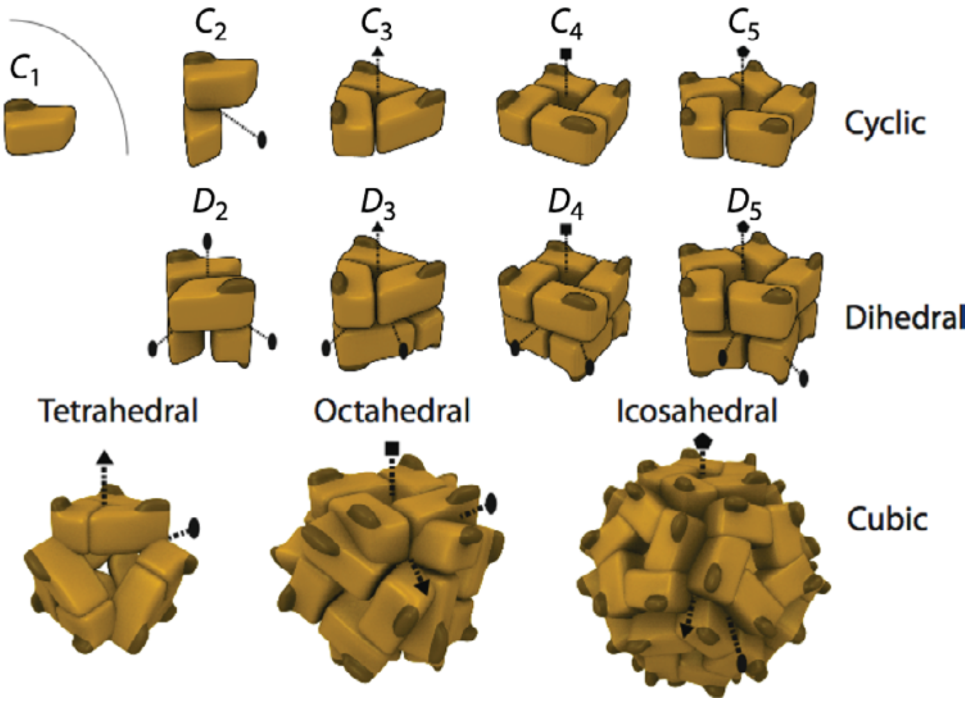


Fig. 1. Symmetry of oligomeric proteins. An oligomeric protein with *n* identical subunits may have the symmetries of the cyclic C_n point group (top row), one with *2n* subunits, the symmetries of the dihedral D_n point group (middle row); cubic symmetries (bottom row) require the protein to have 12, 24 or 60 identical subunits. Symmetry axes of different types are marked as dotted lines. Courtesy of E. Lévy (Cambridge, UK).

(Fig. 1). Oligomers that display the symmetries of a cyclic C_n group have an *n*-fold axis: their subunits are related by $360°/n$ rotations. The dihedral D_m groups require an even number of subunits, $n=2m$; they possess an *m*2-fold axis and *m*2-fold axes orthogonal to it. The *T* (tetrahedral) cubic point group has non-orthogonal twofold and threefold axes; in addition, the *O* (octahedral) point group has fourfold axes, and the *I* (icosahedral) point group, fivefold axes.

Symmetry is a general property of oligomeric proteins (Goodsell & Olson, 2000). The most common is C_2 in homodimers, but in larger oligomers, dihedral symmetry is much more frequent than cyclic symmetry, for soluble proteins at least. Thus, D_2 tetramers are more common than C_4, and D_3 hexamers are more common than C_6. In contrast to soluble proteins, membrane proteins do not normally display dihedral symmetry, incompatible with the polarity of biological membranes, but they often have cyclic symmetry. Examples are the C_3 trimer of bacteriorhodopsin, the C_4 tetramer of the potassium channel and the C_5 pentameric acetylcholine receptor. Cubic symmetry requires *n* to be a multiple of 12 in the *T* point group, 24 in *O* and 60 in *I*. As a consequence, it is present only in large oligomers, and the best-documented example is the icosahedral symmetry of the viral capsids discussed below.

2.3 Quaternary structure and the Protein Data Bank

The Protein Data Bank (PDB; Berman *et al.* 2000) stores atomic coordinates issued from X-ray and NMR studies. In April 2008, the PDB contained more than 50 000 entries describing the atomic structure of some 20 000 different proteins. It should be the natural place to look for their

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

QS, yet deriving a QS from the information in a PDB entry is cumbersome and sometimes misleading. The reason is intrinsic to crystallography: in a protein crystal, inter-molecular contacts coexist with the subunit contacts that define the QS, and distinguishing one from the other is sometimes not straightforward. Algorithms specifically developed for this purpose have been reviewed by Poupon & Janin (in press). The problem does not arise for NMR structures, which are determined in solution, but few NMR studies address oligomeric proteins or protein–protein complexes, due to their larger size and symmetry that creates ambiguities when assigning resonances.

By convention, a crystallographic PDB entry reports atomic coordinates for the crystal asymmetric unit (ASU), rather than the molecular assembly in solution, which the PDB defines as the biomolecule. There is no simple relation between the ASU and the biomolecule: a monomeric protein can yield crystals with two or more chains in the ASU, an oligomeric protein, crystals with only one chain, in which case its symmetry must be a crystal symmetry. The QS is often not mentioned as such in a PDB entry, and when the word 'dimer' appears, the protein needs not be a dimer in solution. Since 1999, most PDB entries contain two records that define the biomolecule. REMARK 300 relates its subunit composition to the content of the ASU; REMARK 350 cites the matrices needed to build it from the ASU. Thus, if a homodimeric protein crystallizes with a monomer in the ASU, REMARK 300 will mention one chain and REMARK 350 two matrices. But if there is a dimer in the ASU, REMARK 300 will cite two chains, and REMARK 350, only the identity matrix.

Converting this information into a QS requires some effort, but several databases offer that service and give access to the atomic coordinates of the biomolecule: Biounit, ProtBuD and 3 D-Complex (described in Section 2.4). Biounit, accessible through the PDB interface at the Research Collaboratory for Structural Bioinformatics (RCSB; Rutgers University, New Jersey), relies on REMARK 300/350 or on supporting information from the authors if the records are absent. ProtBuD (Protein Biological Unit Database; Xu *et al.* 2006) reports the QS of the biomolecule in both the PDB and the PQS database. Probable Quaternary Structure (PQS; Henrick & Thornton, 1998), PITA (Protein InTerfaces and Assemblies; Ponstingl *et al.* 2003) and PISA (Protein Interfaces, Surfaces and Assemblies; Krissinel & Henrick, 2007), implement the approach of the problem developed at the European Bioinformatics Institute (EBI-EMBL, Hinxton, UK). It is based on the geometric and physical chemical properties of the interfaces between molecules, and ignores the information in the header of a PDB entry, although the two agree in 82% of the cases (Xu *et al.* 2006). PQS and PITA apply crystal symmetries to the molecules in the ASU, generate neighbors and score each pairwise interface on the basis of the buried area, plus a solvation energy term in PQS or a statistical potential in PITA. The QS is then iteratively built by retaining the interfaces that achieve high scores. In PISA, the iterative construction is replaced by a graph exploration that surveys all the assemblies that can be formed in the crystal. PISA handles non-protein components, and it may detect assemblies missed by PQS or PITA. Given a PDB code or a set of atomic coordinates, the user interfaces of all three servers return information on the pairwise interfaces and the assemblies that pass their respective criteria, and they allow downloading their atomic coordinates.

2.4 Mining the biochemical literature

The QS information in the PDB is not documented and may not be updated when new data become available. It should therefore be completed, and possibly corrected, by surveying the

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

biochemical literature and identifying data that concern the protein assembly in solution. The analysis of the interfaces in protein–protein complexes and oligomeric proteins described in Section 4 below has been performed on curated sets that were assembled by manual surveys, and represent only a small fraction of the PDB. Recently, Lévy (2007) carried out a large-scale literature search with keywords related to the QS and to methods for molecular weight determination. He was able to assign the QS of more than 3000 proteins, and cover about one-quarter of the PDB by extending the assignment to close homologs. The agreement with the curated datasets is nearly perfect, but the annotated QS disagrees with the PDB biomolecule in about 15% of the entries, and in up to 27% of the proteins with non-redundant sequences.

The results of the search are accessible through the PiQSi (Protein Quaternary Structure Investigation; Lévy, 2007) database, which is derived from the 3D Complex database (Lévy *et al.* 2006), and interlinked with it. Like Biounit, 3D Complex relies on the information in the PDB entries, but its graph description of the QS and its hierarchic structure are original. The QS hierarchy of 3D Complex, shared with PiQSi and inspired of the domain hierarchy in SCOP (Murzin *et al.* 1995), has a top level of 'topologies' that depend on the number of subunits, the symmetry and the pattern of contacts in the molecular assembly. Below, it has 'families' that represent evolutionary relationships, and QSx 'classes' in which x is the sequence identity between equivalent chains in related assemblies. PiQSi, which initially contained about 10 000 entries, is being updated to cover the whole PDB. When a PDB code or a protein sequence is entered, the interface displays the protein QS as a graph, and cites the MedLine ID code of the references used to annotate it (Fig. 2). A tag indicates whether the biomolecule in the PDB is thought to be correct, incorrect or uncertain, and points to a comment that supports the annotator's opinion. PiQSi has another very valuable feature: users can submit new annotations that will be processed by the curators and eventually propagated to the database.

3. Tools to study macromolecular interfaces

3.1 Geometry and the definition of interfaces

3.1.1 The buried surface model

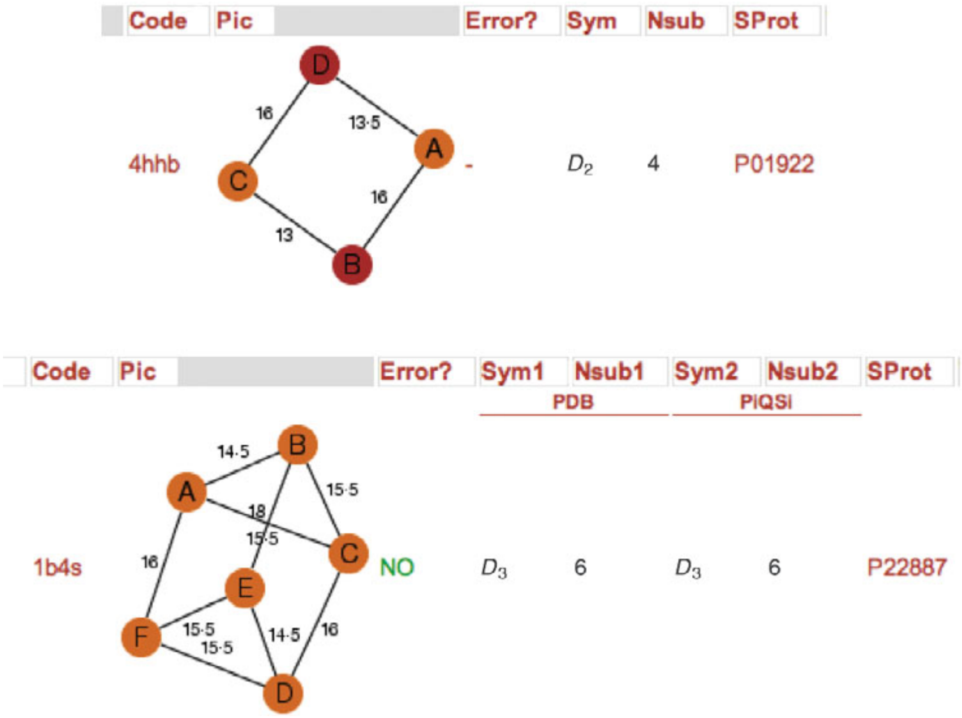
Given the atomic structure of a macromolecular assembly, defining the interface between two components A and B may be viewed as a problem of geometry in space. The simplest definition is based on distance: the AB interface is the set of atoms or chemical groups i of A and j of B, which satisfy the condition $d_{ij} < d_0$. In most implementations, d_0 depends on the atomic or group radii r_i and r_j, and on a cutoff value r_0 in the range 0·5–2 Å:

$$d_{ij} < d_0 = r_i + r_j + r_0 \qquad (1)$$

The interface can also be defined as the surface buried between the two components (Chothia & Janin, 1975). Given the solvent accessible surfaces of A, B and the AB pair, any point of the accessible surface of A or B that does not belong to the accessible surface of AB is part of the interface. The solvent accessible surface is at one probe radius ($r_W = 1{\cdot}4$–$1{\cdot}5$ Å for a water probe) of the molecular surface, and its construction may implement the rolling sphere algorithm of Lee & Richards (1971), or an analytical algorithm. The buried surface area, or interface area, that represents the interface size, is computed as:

$$\mathrm{BSA} = \mathrm{ASA}_A + \mathrm{ASA}_B - \mathrm{ASA}_{AB} \qquad (2)$$

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press



Code	Pic	Error?	Sym	Nsub	SProt
4hhb		-	D_2	4	P01922

Code	Pic	Error?	Sym1	Nsub1	Sym2	Nsub2	SProt
			PDB		PiQSi		
1b4s		NO	D_3	6	D_3	6	P22887

Fig. 2. Graph representation of the protein QS. The PiQSi database (Lévy, 2007) reports the QS of proteins in the PDB, checked against the literature. 4hhb (top): Human hemoglobin is shown as a graph where the nodes are the α and β subunits (in two different colors), and the edges are the subunit contacts in the tetramer; the label next to each edge is the number of residues implicated; the tetramer has (approximate) D_2 symmetry. 1b4s (bottom): The nucleoside diphosphate kinase of *Dictyostelium discoideum* is a homohexamer with D_3 symmetry; 'NO' in the 'Error?' column indicates that the literature agrees with the QS described in the PDB entry.

where $\mathrm{ASA_A}$, $\mathrm{ASA_B}$ and $\mathrm{ASA_{AB}}$ are the accessible surface areas of A, B and the AB pair. In that model, the atoms and residues that lose ASA in AB are part of the interface. The BSA is essentially proportional to their number, and also to the number of atom pairs that satisfy Eq. (1) with $r_0 = 2r_\mathrm{W}$.

Several servers, including PP at University College London, and PISA at the European Bioinformatics Institute (see Section 2.3), return the areas of the solvent accessible (ASA) and of the buried surfaces (BSA) of individual polypeptide chains, residues and atoms, when atomic coordinates are entered. These servers, and many publications, quote a BSA 'per subunit', that is BSA/2 in Eq. (2). This assumes implicitly that A and B bury the same surface area, which is exact only for interfaces with C_2 symmetry. The difference between the contributions of the two partners to the BSA depends on the shape of the interface. In most cases, it is only a few percents, but it may exceed 10% when a convex surface is in contact with a concave surface; examples are protease inhibitors binding to proteases (Lo Conte *et al.* 1999) and RNA binding to proteins (Bahadur *et al.* 2008).

3.1.2 The Voronoi model

An alternative definition of interfaces is based on the Voronoi diagram and a related construction, the alpha-complex. The Voronoi diagram associates to each atom its Voronoi cell, the

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

convex polyhedron that contains all points of space closer to that atom than to any other atom. Its first application to proteins was developed by Richards (1974) to evaluate the volume occupied by individual atoms or chemical groups. Because it does not account for the different sizes of the chemical groups, the Voronoi diagram is at present replaced in most applications by the closely related power diagram (Gellatly & Finney, 1982; Aurenhammer, 1987). Instead of the Euclidean distance $|\mathbf{ax}|$ between a point $\mathbf{x}$ and an atom centered in $\mathbf{a}$, this diagram uses the power distance $p(\mathbf{x})$ of $\mathbf{x}$ with respect to the ball of radius r that represents the atom or chemical group:

$$p(\mathbf{x}) = |\mathbf{ax}|^2 - r^2 \tag{3}$$

The Voronoi cell of an atom then comprises all points of space that have a power distance to that atom less than to any other atom. Its facets belong to the radical plane, which contains the intersection of the spheres if they do intersect.

The Voronoi (or power) diagram offers a natural definition of contacts: two atoms are in contact if and only if their Voronoi cells share a facet. Labeling molecules A and B with different colors, the set of 'bicolor' facets shared by atoms of A and B forms the AB Voronoi interface (Fig. 3). However, the atoms on the molecular surface all have unbounded or poorly defined Voronoi cells, a problem that may be circumvented by placing solvent molecules on the surface (Soyer *et al.* 2000; Dupuis *et al.* 2005), or solved in a mathematically more rigorous manner by using an extension of the power diagram. This extension, called the alpha complex (Edelsbrunner & Mucke, 1994), makes use of a remarkable property of the power diagram: it is invariant when a constant α is added to all squared radii, r^2, in Eq. (3). For a given value of α, the alpha complex is built like a power diagram, except that one restricts the Voronoi cell of each atom to its associated ball and seeks intersections between these restricted regions. Thus, a facet between two atoms is not part of the alpha complex if the associated balls do not intersect, or if the facet lies outside the intersection.

Applications of the alpha complex to macromolecules have been reviewed by Poupon (2004). The interfaces defined in this manner may still contain a few unbounded facets that Edelbrunner and collaborators (Ban *et al.* 2004; Headd *et al.* 2007) removed through an elaborate iterative retraction procedure. The algorithm of Cazals *et al.* (2006), implemented on the Intervor server, does it on purely geometric criteria. Both procedures have been applied to a set of protein–protein interfaces that had been previously analyzed with the buried surface approach by Chakrabarti & Janin (2002). The Voronoi interface area, calculated as the sum of the areas of the bicolor facets, correlates linearly with the BSA in both cases, but the correlation is much better with the procedure of Cazals *et al.* (2006) than that of Ban *et al.* (2004) ($R^2 = 0{\cdot}98$ *vs.* $0{\cdot}85$). A remarkable result is that about 13% of the atoms that share bicolor facets do not contribute to the BSA, mostly because they are not solvent accessible in the free molecules. Thus, the Voronoi interface comprises significantly more atoms, and especially more main chain atoms, than the buried surface.

3.2 Topology: modules, patches, core, rim and segments

The topology of an interface represents the various ways in which it can be split. The contacts between macromolecules often implicate several regions on the surface of each partner, and these regions have been given different names by different authors: interaction sites (Jones & Thornton, 2000), recognition patches (Chakrabarti & Janin, 2002), interface modules (Reichmann *et al.* 2005, 2007). They often belong to distinct structural domains that may constitute autonomous units in the proteins, each with a specific interaction pattern (Copley *et al.*

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

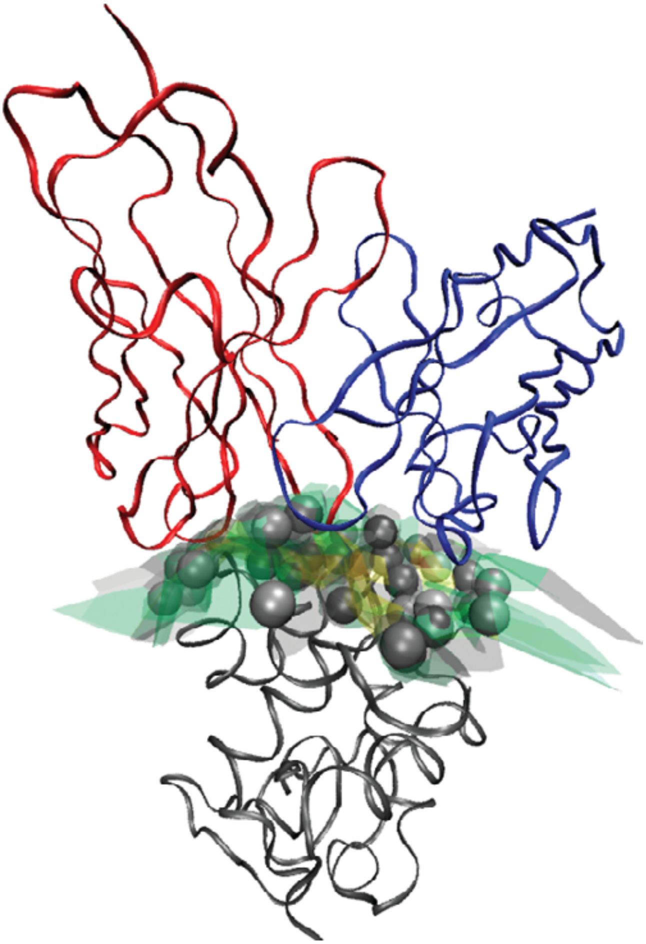


Fig. 3. The Voronoi model of a protein–protein interface. The interface between hen egg white lysozyme (gray ribbon) and the Fv fragment of antibody D1.3 (heavy chain in red, light chain in blue) is drawn as a set of Voronoi facets that belong to the alpha-complex for $\alpha = 0$. Unbounded facets have been removed by the procedure of Cazals *et al.* (2006). The interface is heavily hydrated with the water molecules drawn as gray spheres. PDB entry 1vfb (Bhat *et al.* 1994). Courtesy of F. Cazals (Sophia-Antipolis, France).

2002). Classical examples are the immunoglobulin domains or the Sarc homology SH2 and SH3 domains, but databases such as InterPro (Apweiler *et al.* 2001), PFAM and its derivative iPFAM (Finn *et al.* 2005) list thousands more. The topology of an interface may also be defined biochemically. The interface modules of Reichmann *et al.* (2005) are obtained by grouping residues on the basis of the cooperative interactions observed in double mutant cycle experiments: the interactions are high within a module, and negligible between modules. In the systems studied by Reichmann *et al.* (2005), clustering the residues by distance yields the same grouping as the cooperative interactions. This suggests that the modules should resemble the 'hot regions' of Keskin *et al.* (2005), which group adjacent conserved hotspots from alanine-scanning experiments (see Section 4.1.3).

The recognition patches of Chakrabarti & Janin (2002) are identified by applying to the interface atoms the average linkage method, a geometric clustering algorithm that relies on a threshold distance d_m. The algorithm is run separately on each component protein, and therefore

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

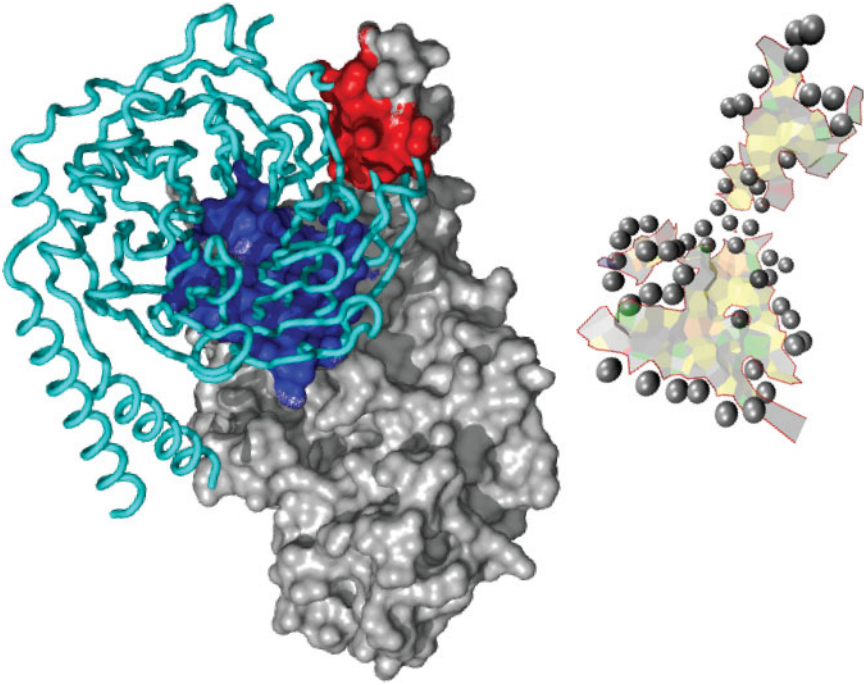


Fig. 4. Two views of the topology of a protein–protein interface. The heterotrimeric G-protein transducin (1got; Lambright *et al.* 1996) is drawn with main chain tubes in cyan for the Gβ and Gγ subunits, and a gray surface for the Gα subunit. The geometric clustering procedure of Charkrabarti & Janin (2002) splits the region of the Gα surface in contact with G$\beta\gamma$ into two recognition patches painted blue and red. The red surface belongs to the N-terminal helix of Gα, which is disordered in free Gα. On the right, the Voronoi model of the interface (Cazals *et al.* 2006) is seen to give the same description of its topology: the two connected components correspond to the two patches, each surrounded by hydration water.

one component may contain more clusters than the other. However, in the tests performed by Chakrabarti & Janin (2002) on a set of 70 protein–protein complexes, the difference was never more than one with the default value $d_m = 15$ Å, half the maximum distance between atoms in the set, and it could be brought to zero by adjusting d_m to values in the range 15–20 Å. Thus, the recognition patches generally occur in pairs, one on each component protein. The 70 complexes contain 46 single-patch interfaces (with one patch on each component), and 24 that are multi-patch.

Jones & Thornton (1997) used a related procedure in a study aiming to predict interaction sites from amino acid sequences. *n* being the number of residues in an observed interface, they defined surface patches as sets of *n* solvent-accessible residues surrounding a central accessible residue, and compared the values of six different parameters in the patches that overlapped with the actual interface and those that did not. Three of the parameters were physical–chemical: a solvation potential, a residue interface propensity and hydrophobicity. The other three were geometric: the mean ASA of the residues in the patch, a planarity and a protrusion index. Planarity was measured by the root mean squared deviation of all the patch atoms from the least squares plane through the atoms, protrusion, by fitting an ellipsoid to the protein or domain (Thornton *et al.* 1986).

The Voronoi model offers an alternative definition of the topology. The set of facets that form the Voronoi interface may be split into subsets of facets that have edges in common (Fig. 4). The

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

subsets, named connected components by Cazals *et al.* (2006), correlate well with the recognition patches identified by the clustering algorithm in the 70 interfaces of Chakrabarti & Janin (2002), and because their definition does not depend on a distance cutoff, they may represent a more robust description of the interface topology.

Another approach distinguishes between the regions of an interface that are fully buried and those that remain partly accessible to solvent. In the macromolecular assemblies considered here, only a minority (20–45%) of the interface atoms have zero ASA, and they tend to belong to hydrophobic residues. Chakrabarti & Janin (2002) define the interface core as the set of residues that contain atoms buried at the interface; the remaining interface residues form the rim. Section 4.1.3 relates the core/rim topology to the 'O ring' model of protein–protein interfaces put forward by Bogan & Thorn (1998) based on site-directed mutagenesis data. These data, collected in the ASEdb database (Thorn & Bogan, 2001), suggest that the 'hotspot' residues, those that have a large effect on the stability of a complex when mutated, tend to be surrounded by energetically less important residues that occlude bulk solvent from them. The former should therefore belong to the core of the interface, and the latter, to its rim. Reichmann *et al.* (2005) and Keskin *et al.* (2005) have reached similar conclusions from different starting points.

Interfaces can also be split into segments along the polypeptide chain. Jones & Thornton (1996) define interface segments as stretches of residues that start and end with an interface residue. A segment may contain intervening non-interface residues, but only in stretches of n residues or less; $n=5$ in their implementation, $n=4$ in that of Pal *et al.* (2007) and in the data presented in Section 4. The number of segments, typically 1–8 in protein–protein complexes, and that of interface residues per segment, typically 1–10, describe how these residues are distributed along the polypeptide chain. Another point of interest is their distribution into secondary structure elements (Guharoy & Chakrabarti, 2007).

3.3 Atomic packing, cavities and shape complementarity

The Voronoi diagram allows making a precise estimation of the atomic packing. The Voronoi cell of an atom buried inside a protein represents the space it occupies, and its volume is essentially determined by its chemical nature (Richards, 1974). The volumes occupied by atoms buried inside proteins are the same, or possibly slightly smaller, as in crystals of small molecules or peptides (Harpaz *et al.* 1994; Pontius *et al.* 1996; Tsai *et al.* 1999; Tsai & Gerstein, 2002). Therefore, the interior of proteins must be tightly packed in spite of the occasional cavities. The Voronoi volumes of surface atoms, obtained by modelling the solvent around the protein, are about 7% larger (Gerstein *et al.* 1995; Gerstein & Chothia, 1996), which suggests that the packing is less tight on the protein surface.

Voronoi cells can also be constructed for atoms buried at interfaces. Those are relatively few, but the construction can make use of the crystallographic solvent molecules reported in high-resolution X-ray structures, in which case the volume measurement can be made on a majority of the interface atoms. Atoms at the interfaces of protein–protein complexes (Lo Conte *et al.* 1999) and of oligomeric proteins (Ponstingl *et al.* 2005) have been shown in this way to occupy the same volume as inside proteins to within 1–2%. This implies that they are close-packed, and therefore, that the molecular surfaces buried in the contact have complementary shapes. However, the fraction of the interface atoms that are buried is itself a parameter that depends on the quality of the packing. A poor fit of the molecular surfaces creates cavities (Hubbard &

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Argos, 1994), and results in an interface that buries comparatively few atoms, yet these atoms may still be close-packed.

Several other estimates of the atomic packing at interfaces have been proposed over the years. The first was the shape complementarity index S_c of Lawrence & Colman (1993). Given the coordinates of the AB complex, S_c is evaluated by first using the program MS of Connolly (1983) to define grid points of the molecular surface of A and B, and calculating a local index $S(\mathbf{x})$ at all grid points $\mathbf{x}$:

$$S(\mathbf{x}) = \cos u \exp(-wd^2) \tag{4}$$

where d is the distance of $\mathbf{x}$ on A to the closest grid point on B, u is the angle of the normal vectors to the surfaces at these two points, w is a weight factor adjusted to 0·5 Å^{-2}. S_c is the median value of $S(\mathbf{x})$ after removing grid points on the edge of the contact region that yield abnormal low values. Protein–protein complexes and oligomeric proteins typically have S_c values near 0·70 (Lawrence & Colman, 1993; Bahadur *et al.* 2004).

Another approach to atomic packing is to detect cavities and measure their volume. Program SURFNET (Laskowski, 1995) does that by fitting spheres in between the two molecular surfaces; spheres in contact with atoms at the edge of the interface are removed if their radius exceeds 10 Å. The volume of the union of the spheres is the gap volume V_{gap}, and the ratio to the buried surface area, the gap volume index:

$$I_{\text{gap}} = V_{\text{gap}}/\text{BSA} \tag{5}$$

Typical values of I_{gap} in complexes and oligomeric proteins are in the range 2–3 Å (Jones & Thornton, 1996). The PP server at University College London allows calculating V_{gap} and I_{gap} from input atomic coordinates. Because the larger spheres tend to be on the edge, the gap volume index may take inconsistent values when the interface is small, or split into many patches that have a high proportion of edge atoms. This remark also applies to the edge removal step in the S_c index calculation.

The 'local' L_D and 'global' G_D indexes of Bahadur *et al.* (2004) are less sensitive to edge effects. L_D is the mean value over i of the local density of the interface atoms, estimated as the number n_i of those that are within a given distance D of interface atom i. With $D = 12$ Å, L_D is in the range 40–45 for most complexes and oligomeric proteins. G_D, which also represents a surface density, is estimated as the ratio N/A_{ellips}, where N is the number of the interface atoms and A_{ellips} is the area of the equatorial section of their ellipsoid of inertia. Typical values of G_D are near 1·3 atom Å^{-2} (Bahadur *et al.* 2004). The ProFace server of the Bose Institute, Calcutta (Saha *et al.* 2006), returns values of several interface parameters including L_D when coordinates are uploaded.

3.4 Chemical and physical–chemical properties

3.4.1 Chemical and amino acid compositions

The chemical composition of an interface is commonly characterized by the relative abundance of the different types of atoms or amino acid residues. Atom types are often defined as non-polar/polar/charged, or main chain/side chains. The latter distinction is useful because the protein main chain contributes significantly to most macromolecular interfaces, and biochemical approaches based on site directed mutagenesis tend to ignore that contribution. In addition to protein atoms, the interface may comprise water molecules (see Section 3.4.3).

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

A composition may be evaluated by counting interface atoms or residues, or by measuring their contribution to the BSA. The two definitions are nearly equivalent for atoms, but with residues, the area-based composition reflects their size as well as their abundance. In either case, the fractions may be converted into propensities:

$$p_i = \ln\left(f_i/f_i^{\circ}\right) \qquad (6)$$

where f_i is the number or area fraction of type i at the interface, f_i°, the corresponding number in a reference set that can be the whole protein, its interior or its surface. If the reference is the protein surface, f_i° may be either a fraction of the number of solvent-accessible atoms or residues, or a fraction of their contribution to the ASA. Then, $p_i > 0$ implies that interfaces are enriched in atoms or residues of type i relative to the protein surface in contact with the solvent; $p_i < 0$ implies that the interfaces are depleted in type i. An interface commonly comprises only a few tens of residues, and propensities calculated on individual interfaces or on small sets with 20 residue types have poor statistics. Grouping types with similar properties may improve the statistics, but this will mask at least one interesting feature: Arg and Lys are usually counted as similar, yet protein–protein interfaces are depleted in Lys, but not Arg.

3.4.2 Hydrophobicity

Hydrophilicity/hydrophobicity is an important physical chemical property of the protein surface, and possibly the one most relevant to its propensity to interact with a ligand. It can be estimated at the atomic level: aliphatic and aromatic groups in proteins are non-polar, and therefore hydrophobic; O- and N-containing groups are polar and hydrophilic. When they interact with a ligand, all these groups become partly or fully dehydrated, and groups of the ligand also, allowing solvent molecules to gain degrees of freedom. Their entropy increases, causing the hydrophobic effect to favor association. The system enthalpy also changes, and it can do so in either direction, because van der Waals and polar interactions made by surface groups with the solvent are replaced by protein–ligand interactions. The whole process can be viewed as a transfer of the ligand from water to the protein environment. With hydrocarbons, which make similar van der Waals interactions with water and other solvents, and no polar interaction, the free energy of transfer scales linearly with the BSA. Calibrations based on their relative solubility in water and organic solvents yield a slope of 20–30 cal mol^{-1} Å^{-2} (Chothia, 1974; Vajda *et al.* 1995). Site-directed mutagenesis experiments discussed in Section 4.1.3 suggest that this correctly represents the contribution of the hydrophobic effect to protein–protein interaction, although the contribution of individual hydrophobic groups may depend on the structural context.

3.4.3 Polar interactions and hydration

Electrostatic interactions implicate polar groups containing oxygen or nitrogen atoms. These atoms bear a partial or a full charge, and they form H bonds, much stronger than van der Waals interactions. Electrostatics plays an essential role in all processes involving proteins, DNA or RNA, but its evaluation requires elaborate calculations that are highly sensitive to the solvent model (Sheinerman *et al.* 2000). The H bonds between protein groups, or between the protein and a ligand, have an energy similar to those made with water, and the balance depends on the details of their geometry and their environment. Program HBPLUS (McDonald & Thornton, 1994) is commonly used to detect H bonds in protein structures; with a donor–acceptor distance

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

cutoff of 3·9 Å, the default value in its implementation by the PP server of University College London, it retains H bonds that may be much weaker than that those between water molecules in the liquid, where the oxygen–oxygen distance is 2·8 Å. H bonding is mostly a dipole–dipole interaction, but electrostatics also involves long-range interactions between charged groups, and weaker interactions involving aromatic groups or water bridges. Although efficiently screened by salts under most conditions, long-range electrostatic interactions may play an active part in the kinetics of association (Schreiber *et al.* 2006).

The representation of electrostatics in force fields is not accurate enough to give reliable values of the energy balance of these interactions in a water environment. In practice, all-atom force fields perform best if they are calibrated on actual protein structures. An example is the Rosetta force field, which has yielded excellent models of several proteins and protein–protein complexes (Bradley *et al.* 2005; Schueler-Furman *et al.* 2005). Energy refinement with an all-atom force field is computationally expensive, albeit well suited for distributed computing (Das *et al.* 2007). It is usually left for late steps of structure predictions, and earlier steps rely instead on knowledge-based statistical potentials, of which TASSER (Zhang *et al.* 2005) is an example. Knowledge-based potentials, originally developed to model protein folding, are now commonly used for protein design (Poole & Ranganathan, 2006). They can also be applied to protein interactions (Keskin *et al.* 1998), but the tendency in recent years has been to develop potentials that are specifically designed for that purpose (Mintseris *et al.* 2007), and to complement the statistical approach with machine-learning procedures that can handle parameters not easily converted into potentials (Block *et al.* 2006; Zhu *et al.* 2006; Ofran & Rost, 2007; Bernauer *et al.* 2008).

The interface between two proteins, or between a protein and a nucleic acid in a complex, is never completely dehydrated. This can be seen in high-resolution X-ray structures that report water molecules: many are located at molecular interfaces. Rodier *et al.* (2005) consider as interface waters all crystallographic waters located within 4·5 Å of atoms of both subunits; nearly all are within H-bond distance of at least one protein polar atom. Their number increases with the interface size, but also with the resolution of the X-ray structure, which strongly suggests that solvent is under-reported in most studies.

3.5 Conformation changes

Conformation changes, which occur frequently when a protein interacts with a ligand, can be assessed in cases where X-ray or NMR structures are available for both the free and the liganded protein. The ligand can be a small molecule, a nucleic acid or another protein, and the change may affect just a few atoms in the region of contact, or the whole protein. Its amplitude can be estimated by performing a least-square superposition of the free and liganded protein, and measuring the root-mean-square displacement (RMSD) of the Cα atoms. As the RMSD tends to increase with the protein size, it may be useful to normalize it (Carugo & Pongor, 2001), or use instead a comparison of the Cα–Cα distances in the free and liganded structures.

The RMSD ranges between 0·2 and 8 Å in a set of 124 protein–protein complexes for which the free component structures are available (Hwang *et al.* 2008). Half of the complexes in that set have an RMSD <1 Å, comparable to what is obtained when two different X-ray structures of the same protein are superimposed. They display only side chain rotations and localized main chain movements, and can be assumed to form by rigid-body association. By contrast, the complexes with a large RMSD display movements of all sorts: of surface loops, of secondary structure

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

elements and, in some cases, of whole domains. The conformation observed in complexes may preexist the interaction, or reflect an induced fit that follows the encounter of the two molecules (Grünberg *et al.* 2006), but in either case, the conformation change is an essential element of the recognition process.

3.6 Conservation in evolution

Proteins with closely related sequences usually have the same QS, but this breaks down at a certain level of divergence. Hemoglobin is a tetramer in mammals, but not invertebrates, and sea lamprey, a vertebrate, has a monomeric hemoglobin with 31% sequence identity to human. This example shows that QS is much less conserved than the secondary or tertiary structure, which is the same in all globins. Lévy *et al.* (2008) find that, at a 30–40% sequence identity level, 30% of the proteins in PiQSi have a different QS, and that the QS changes in half of the homologs with less than 30% identity. Thus, homology does not reliably predict the QS at levels of identity where the fold is certainly maintained, and this may be why standard tools for homology modelling generally ignore the QS.

An earlier report by Aloy *et al.* (2003) finds that the interaction between domains is almost invariably conserved above 30% sequence identity. This finding is at the origin of the InterPreTS (Interaction Prediction through Tertiary Structure; Aloy & Russell, 2003) procedure. It uses a library of Pfam domains known to interact, called 3DID (Stein *et al.* 2005). If two proteins contain sequences related to a domain pair in 3DID, the sequences are aligned with the closest homologs, and the interaction modelled on that in the pair. InterPreTS was used to carry out structural predictions on 102 protein complexes in yeast, yielding at least partial models of half of the complexes (Aloy *et al.* 2004). At low levels of sequence identity, threading methods may complement homology in QS predictions. The M-TASSER procedure (Grimm *et al.* 2006) was designed for predicting homodimers and, like InterPreTS, it uses a library of templates. A query sequence is threaded with TASSER (Zhang *et al.* 2005) to generate models of the monomers that are structurally aligned on the dimers in the library, and then refined. The structural alignment identified a correct template in about half of the sequences of a test set, and the refined models had an average RMSD of 5·9 Å relative to the native structures (Chen & Skolnick, 2008).

The divergence of the QS during protein evolution may also be evaluated at the residue level. A conserved QS implies conserved interfaces, and therefore, a selection pressure on the interface residues. Given the aligned sequences of N homologous proteins, the variability at sequence position i is commonly measured by the Shannon entropy (Sander & Schneider, 1993; Elcock & McCammon, 2001; Caffrey *et al.* 2004; Guharoy & Chakrabarti, 2005):

$$s(i) = -\sum_{k=1,T} p_k \ln p_k \tag{7}$$

p_k is the fraction of the sequences that have in i a residue of type k; $s(i)=0$ at fully conserved positions, $s(i)=\ln T$ at totally divergent positions. The number T of types can be 20, or less if types with similar properties (e.g. hydrophobic) are merged (Elcock & McCammon, 2001). The entropy depends on the choice of the aligned sequences, their number and their diversity, but it can be normalized to the average value in the polypeptide chain in order to limit that dependence.

Valdar & Thornton (2001) and Armon *et al.* (2001) use a similar approach with more elaborate measures of the evolutionary divergence that take into account observed mutation rates, in order

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

to predict binding sites. The procedure of Valdar & Thornton (2001) is implemented in the Scorecons server of the European Bioinformatics Institute, and that of Armon *et al.* (2001) in the ConSurf server of Tel Aviv University. The Evolutionary Trace procedure (Lichtarge *et al.* 1996; Res & Lichtarge, 2005; Mihalek *et al.* 2006), also designed to predict binding sites, builds phylogenetic trees from sets of homologous sequences. In addition to residue conservation, pairwise interactions between proteins may be identified by searching for the co-evolution of their amino acid sequences, but the accuracy is poor unless the structure of a homologous complex is available (Pazos & Valencia, 2002; Valencia & Pazos, 2002; Mintseris & Weng, 2005; Halperin *et al.* 2006; Juan *et al.* 2008).

3.7 Docking predictions

Docking methods aim to model the structure of a complex from that of its components. Protein–protein docking procedures have often been reviewed (Halperin *et al.* 2002; Smith & Sternberg, 2002; Marshall & Vakser, 2005; Gray, 2006; Wiehe *et al.* 2007). Several are implemented as Web servers (Table 1). Their performance can be tested on benchmark sets of protein–protein complexes (Gao *et al.* 2007; Hwang *et al.* 2008). Since 2001, protein–protein docking is assessed by blind predictions in the CAPRI experiment (Critical Assessment of PRedicted Interactions; Janin *et al.* 2003; Janin, 2005; Janin & Wodak, 2007). A CAPRI target is a protein–protein complex that has a known, but still unpublished, experimental structure, and known component structures. CAPRI predictors dock the components and submit models that are assessed against the experimental structure. In six years, 28 target complexes have been submitted to the prediction, which yielded good-quality models of a majority of them. Details of the methods and the results can be found in the evaluation papers (Mendez *et al.* 2003, 2005; Lensink *et al.* 2007), and the special issues of *Proteins: Structure, Function and Bioinformatics* in which they appear.

Most docking procedures perform an exploration step followed by a refinement step. The first step moves one component as a rigid body relative to the other. In the fast Fourier transform (FFT) correlation algorithm, the two protein structures are mapped onto a cubic grid and each grid point weighted to mark its position relative to the molecule. The correlation function of the weights associated to the two proteins is calculated by FFT for all translations of one protein relative to the other, and the calculation is repeated for all orientations. Procedures based on that algorithm are implemented in a number of servers: ClusPro (Comeau *et al.* 2004), SmoothDock, MultiDock (Gabb *et al.* 1997), Gramm-X (Tovchigrechko & Vakser, 2006). Other docking algorithms randomly generate starting positions, and apply heuristic methods to optimize them (May & Zacharias, 2007). In RosettaDock (Gray *et al.* 2003; Wang *et al.* 2007), this is done in two successive steps of Monte Carlo optimization: one that uses a simplified protein model and force field, and the other, explicit atoms and the very successful Rosetta force field originally developed for protein folding. PatchDock (Schneidman-Duhovny *et al.* 2005a) implements an efficient computer vision algorithm in which the protein surfaces are divided into concave, convex and flat patches, and complementary patches are brought together.

The second step of a docking procedure aims to evaluate and rank the models issued from the exploration step; if near-native solutions are identified, they are subjected to further refinement. The scoring functions used for ranking models may include energy, geometric complementarity, propensities and other terms specific to each approach. When external information is available

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

on the complex, for instance from an interface prediction, sequence conservation, or biological data on mutants, it can be used as a constraint during the exploration step, or be incorporated in the scoring function. The HADDOCK program (Dominguez *et al.* 2003; van Dijk *et al.* 2005) handles external data, including NMR data if they exist, as restraints during energy minimization. Conformation changes can also be modelled at that stage. HADDOCK and RosettaDock do it by leaving free some main chain dihedral angles (Chaudhury *et al.* 2007; de Vries *et al.* 2007). Instead, other procedures generate a number of alternative conformations and perform 'cross-docking' on the conformers (Schneidman-Duhovny *et al.* 2005a; Grünberg *et al.* 2006; Krol *et al.* 2007). In either case, flexibility greatly increases the size of the calculation and the number of false positives.

Docking known structures is an appropriate procedure for predicting non-obligate complexes, but this is usually not applicable to oligomeric proteins, so that very few have been CAPRI targets. A remarkable exception has been the envelope glycoprotein of a flavivirus, known to exist both as a dimer in the viral particle, and as a trimer in the form that promotes membrane fusion when the virus infects a cell (Bressanelli *et al.* 2004). In CAPRI, the structure of the trimer had to be predicted from that of the dimer, a difficult task in which only one predictor succeeded. In other systems where an experimental structure of the components is lacking, docking may be performed on models generated *in silico* by homology or threading (Tovchigrechko *et al.* 2002; Aloy *et al.* 2004, 2005; Zhang & Skolnick, 2004).

Assemblies with more than two components can be generated by adding one component at a time, in which case the constraints due to the occupied space are seen to severely constrain the search (Inbar *et al.* 2005). Thus, a structural model of the nuclear pore (Alber *et al.* 2007) could be built by assigning a fold type to domains in the constituent proteins, modelling the domains and assembling them by optimizing a score function under a large number of restraints derived from experiment. The nuclear pore has a C_{16} symmetry that played a major role in that operation. Symmetry has now been incorporated in several of the docking procedures originally designed to model binary complexes (Berchanski *et al.* 2005; Pierce *et al.* 2005; Schneidman-Duhovny *et al.* 2005a, b).

4. The structural basis of macromolecular recognition

A broad classification of the different modes of macromolecular recognition can be based on the time scale on which they occur. The QS of oligomeric proteins mostly represents a permanent association. Protein–protein complexes, which involve partners that have independent existence, are mostly transient, yet they cover a wide range of affinities and lifetimes. Thus, the complex between barnase, a bacterial ribonuclease, and its intracellular inhibitor barstar, has a $K_d \approx 10^{-14}$ M and a half-life of days (Schreiber & Fersht, 1993). In contrast, redox proteins form complexes with $K_d \approx 10^{-6}$ M and half-lives shorter than 1 s, well adapted to their role in electron transfer (Crowley & Carrondo, 2004). The protein–DNA and protein–RNA complexes considered in Section 4.5 cover a range at least as wide. Irrespective of their stability, all these interactions are biologically significant; they play major roles in essential processes of life, and therefore, they are subject to evolutionary selection. In the cell, they compete with all sorts of interactions that result from the random collision of cellular components. Crystal contacts may serve as a model of those: they have the same physical basis as other protein–protein interactions, but they are non-specific and do not undergo evolutionary selection.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Table 2. *Properties of protein–protein interfaces*

Parameter	Protein–protein complexes[a]	Homodimers[b]		Weak dimers[c]	Crystal packing[d]
		Bahadur	Dey		
Number in dataset	70	122	276	19	188
BSA (Å^2)	1910	3900	3700	1620	570, 1510
(S.D.)	(760)	(2200)	(2160)	(670)	(520)
Amino acids per interface	57	104	100	50	48
BSA (Å^2) per amino acid	34	38	37	32	32
Composition (BSA %)					
Non-polar	58	65	65	62	58
Neutral polar	28	23	22	25	25
Charged	14	12	13	13	17
Atomic packing					
f_{bu} (buried atoms %)	34	36	35	28	21
L_D packing index	42	45	43	34	32
S_c complementarity score	0·69	0·70			0·63
R_p propensity score[e]	0·9	4·3	2·1	0·5	−1·1
Chain segments[f]	5·6	3·4	3·2	5·8	6·3
H bonds					
n_{HB} (number per interface)	10	19	18	7	5
BSA per bond (Å^2)	190	210	209	230	280
Water molecules[g]					
Number per interface	20	44			23
Number per 1000 Å^2	10	11			15
Bridging H bonds	6	13			6
Residue conservation[h]					
% in core	55	60			40
s in core and rim	0·65 and 0·80	0·63 and 0·77			0·98 and 0·99

[a] Data from Chakrabarti & Janin (2002) on a subset of the complexes of Lo Conte *et al.* (1999).

[b] Data from Bahadur *et al.* (2003); the set of Dey *et al.* (unpublished data) was derived from the PiQSi database (Lévy, 2007) as described in the text.

[c] Homodimers described in PiQSi (Lévy, 2007) as being in equilibrium with the monomer, based on the literature.

[d] Pairwise interfaces in crystals of monomeric proteins. The first mean BSA value and the S.D. are for the 1320 interfaces in the 152 crystal forms analyzed by Janin & Rodier (1995). All other numbers are for 188 interfaces with BSA >800 Å^2 that were selected among those 1320 interfaces by Bahadur *et al.* (2004).

[e] Score obtained by summing over the whole interface the propensity of individual residues to occur at the interface of homodimers (Bahadur *et al.* 2004).

[f] Number of polypeptide segments per 1000 Å^2 of BSA. A separate dataset of 204 structures has been used for protein–protein complexes (Pal *et al.* 2007).

[g] Data from Rodier *et al.* (2005).

[h] Mean values in subsets that comprise 52 protein components of the complexes (excluding antigen–antibody complexes), 121 homodimers and 102 monomeric proteins in crystal contacts. s is the mean Shannon entropy in aligned sequences. Data from Guharoy & Chakrabarti (2005).

4.1 Protein–protein complexes

4.1.1 Size and topology of the interfaces

Sets of non-obligate protein–protein complexes of known X-ray structure have been assembled by Janin & Chothia (1990), Jones & Thornton (1996), Lo Conte *et al.* (1999) and recently by

Hwang *et al.* (2008). Table 2 reports average properties of the protein–protein interfaces in a set of 70 complexes adapted from Lo Conte *et al.* (1999) by Chakrabarti & Janin (2002). The table compares their interfaces to those of homodimers and to crystal contacts. The interface size measured by the BSA has a mean value of 1910 Å^2. An average of 204 atoms contribute to that BSA; they belong to 57 amino acid residues, that is about 28 residues per component protein. When site-directed mutagenesis experiments are performed on some of these complexes, many of the residues that lose ASA can be changed to Ala without much effect on the dissociation constant; only a fraction, termed 'hotspots', yield large changes in K_d. In the human growth hormone (HGH)/HGH receptor system for instance, 31 interface residues were mutated on the receptor, but only 11 mutants showed a significant loss of affinity for the hormone (Clackson & Wells, 1995). Similar results have been obtained in other systems (Bogan & Thorn, 1998; DeLano, 2002). The discrepancy with the geometric definition of the interfaces is due in part to the limits of the alanine-scanning method: it does not test all the residues (Ala and Gly are almost always omitted), and ignores the interactions made by the main chain, which contributes 20% of the BSA and a majority of the interface H bonds (Lo Conte *et al.* 1999). Nevertheless, the discrepancy is real, and relevant to the topology of the interfaces and the core/rim model discussed below.

Lo Conte *et al.* (1999) noted that all the protein–protein complexes in their set bury more than 1100 Å^2, and that a majority has a BSA in the range of 1200–2000 Å^2. They retained that range as defining 'standard-size' protein–protein interfaces, as opposed to the 'small' interfaces with BSA <1200 Å^2, and the 'large' interfaces with BSA >2000 Å^2. An interface with a BSA of more than 1200 Å^2 implicating about 60 atoms and 16 residues per component, allows the stable, specific association of two proteins. Nearly all antibody–antigen and many enzyme–inhibitor complexes have standard-size interfaces. The barnase–barstar complex mentioned above does, and also several protease–inhibitor complexes with K_d well below 10^{-12} M. Antigen–antibody complexes are not quite as stable: they typically have a nanomolar K_d and half-lives of a few hours (Braden & Poljak, 2000; Sundberg & Mariuzza, 2002).

The recent set of complexes assembled by Hwang *et al.* (2008) confirms and extends these results. The interface size distribution in that set is shown in Fig. 5. The mean BSA and its standard deviation are essentially the same as in Table 2; 7% of the interfaces are small, 55% are standard size and 38% are large. Three complexes in the new set have a BSA near 1000 Å^2, and one buries only 810 Å^2. It occurs in a complex of the Cbl-b ubiquitin ligase with ubiquitin (PDB entry 2oob). Ubiquitin binding promotes the dimerization of the Cbl-b ligase (Peschard *et al.* 2007), which buries additional surface, but several other complexes involving ubiquitin also have a small BSA. They are short-lived like the redox complexes involved in electron transfer, also listed in Fig. 5. Their interfaces are of standard size or smaller (Crowley & Carrondo, 2004); the smallest, with a BSA of 830 Å^2, occurs in a ternary assembly (PDB entry 2 mta). Taken together, the BSA data suggest that the minimum protein surface that must be buried to form a functional complex is of the order of 900 Å^2 and implicates no more than 12 residues on each partner. The ubiquitin ligase/ubiquitin and the redox complexes illustrate biological processes that depend on short-lived interactions associated with a minimal size interface. Such processes are still poorly represented in the PDB in spite of their importance (Noreen & Thornton, 2003b).

With few exceptions, the small and standard-size interfaces have a simple topology: they comprise a single pair of recognition patches as defined by the clustering algorithm of Chakrabarti & Janin (2002), and a single connected component in the Voronoi model of Cazals

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

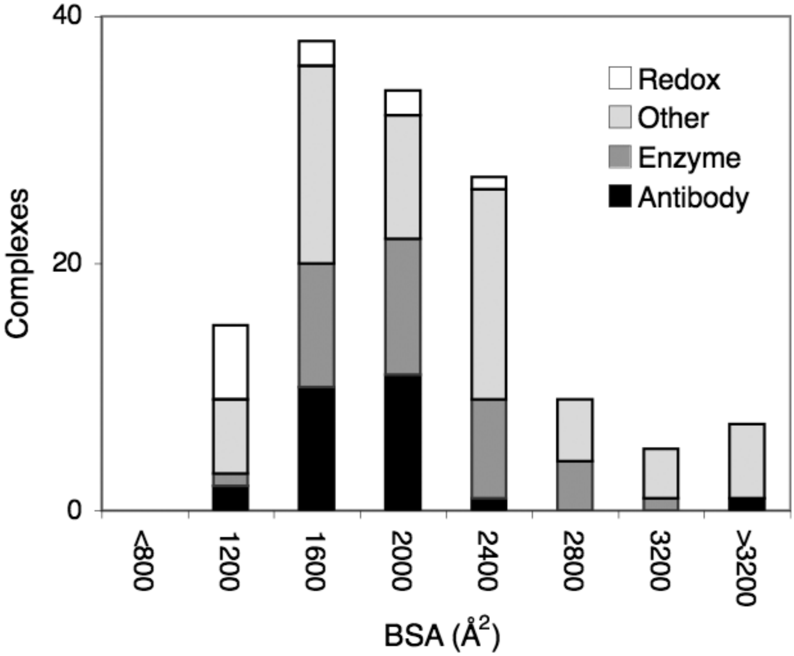


Fig. 5. Interface size in transient protein–protein complexes. Histogram of the BSA in the set of 124 protein–protein complexes assembled by Hwang *et al.* (2008), and the set of 11 redox protein complexes of Crowley & Carrondo (2004). The Hwang set comprises 25 antigen–antibody complexes, 35 enzyme/inhibitor or substrate complexes and 64 complexes of other types. The mean value of the BSA is 1910 Å^2 in that set, and 1290 Å^2 in the redox set.

et al. (2006). The antibody–antigen interface shown in Fig. 3 above is an example of those. Standard-size interfaces are also mostly planar, except in protease–inhibitor complexes where the inhibitor offers a convex surface to the concave active site of the protease (Lo Conte *et al.* 1999). In contrast, the large interfaces usually consist of multiple pairs of recognition patches, at least one of them of standard size (Chakrabarti & Janin, 2002). An example is the Gα/G$\beta\gamma$ interface of transducin (Fig. 4): the blue patch is standard size, and the red patch is small. Individual patches may be planar, but that is unlikely for the whole interface.

The proteins that carry more than one recognition patch on their surface often have the capacity to undergo conformational changes of large amplitude. In transducin, the red patch implicates the N-terminal segment of Gα. This segment is disordered in the free subunit, and forms an α helix in the complex (Lambright *et al.* 1996). Thus, it undergoes a disorder-to-order transition during association. In other systems, the proteins may have domains, each bearing a recognition patch, that move one relative to the other when the complex forms. In the set assembled by Hwang *et al.* (2008), the structure of the free components is known along with that of the complex, and the amplitude of the conformation changes can be estimated by least-square superposition. Figure 6 indicates that the RMSD of the C-alpha atoms tends to increase with the BSA. It is smaller than 1·8 Å in all the complexes with small interfaces, and in 90% of those with standard-size interfaces. This confirms that rigid-body recognition is the preferred mechanism for assembling complexes with standard-size, single-patch interfaces. On the other hand, the RMSD is more than 1·8 Å for 40% of the complexes with large interfaces. Large multi-patch interfaces allow flexible recognition to occur, and we may assume that the additional buried

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

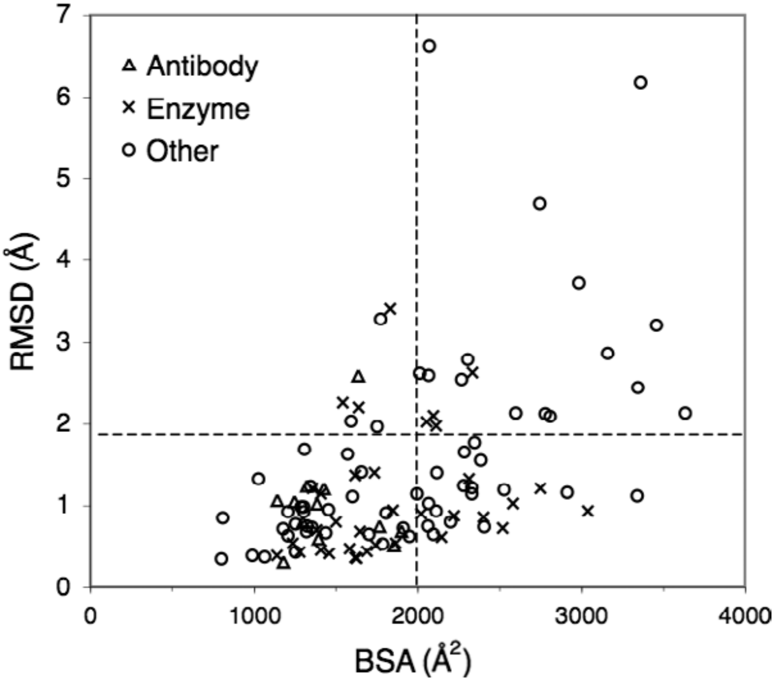


Fig. 6. Interface size and conformation changes. In the set of Hwang *et al.* (2008), the free components of the complexes have a known structure that can be compared to their structure in the complex by rigid-body least-square superposition. The residual RMSD of the C-alpha atoms is plotted against the BSA of each complex. The dotted lines are drawn at RMSD = 1·8 Å and BSA = 2000 Å^2. The categories are the same as in Fig. 5.

surface and atomic interactions pay for the cost of the conformational change (Lo Conte *et al.* 1999; Janin *et al.* 2007).

4.1.2 Composition, packing and hydration

The interfaces of non-obligate complexes have been examined from the viewpoint of their chemical composition. Table 2 shows that the fractional contribution of non-polar (carbon-containing) groups to the BSA is $f_{np} = 58\%$ on average. This value is almost the same as for the ASA of the average monomeric protein ($f_{np} = 57\%$; Miller *et al.* 1987), which reflects the fact that the protein surface involved in an interface remains solvent accessible until the components of the complex meet. Thus, it cannot have very different physical–chemical properties from the remainder of the protein surface. The interfaces of antigen–antibody complexes ($f_{np} = 51\%$) tend to be more polar than the average accessible surface, and those of protease–inhibitor complexes ($f_{np} = 61\%$) are less polar (Lo Conte *et al.* 1999). The interfaces comprise fewer charged groups, and therefore, more neutral polar groups, than the protein surface; the charged groups contribute 14% of the BSA in Table 2 instead of 19% to the ASA (Miller *et al.* 1987). The polar fraction also governs the capacity of interface atoms to form H bonds. On average, transient complex interfaces contain 10 H bonds, about 1 per 190 Å^2 of BSA (Table 2), defined with the geometric criteria of HBPLUS (McDonald & Thornton, 1994). The number of H bonds, n_{HB}, in individual interfaces increases with the BSA, but the correlation is mediocre

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

($R^2=0{\cdot}84$). Moreover, high-resolution X-ray structures tend to display more H bonds, which suggests that the surface density calculated on the whole sample is underestimated.

A large majority of the atoms at protein–protein interfaces is still accessible to the solvent in the complex. The fraction of fully buried interface atoms (zero ASA), f_{bu}, is only 34% on average (Table 2), and it never exceeds 50%. Measurements of the Voronoi volume indicate that the buried interface atoms pack as densely as the atoms of the protein core. This implies that the interfaces are close-packed, a statement that can be extended to 60% of the interface atoms by including the crystallographic water molecules in the construction of the Voronoi diagram. In most cases, the volumes are within 1% of those of the protein core (Lo Conte *et al*. 1999).

Residual water plays an important role in the packing. It is abundant at the interfaces: they contain an average of 20 water molecules, that is 10 water molecules per 1000 Å^2 of BSA (Table 2). This amounts to about 10% of the hydration water present before association, a water molecule covering about 10 Å^2 of the protein surface on average. This estimate of the residual hydration of the interfaces is certainly below the reality, because X-ray structures do not report solvent in a consistent way, and the surface density of interface water increases with the resolution even faster than for H bonds (Rodier *et al*. 2005). Water molecules make many polar interactions and H bond bridges across the interface. The balls representing water in Figs 3 and 4 illustrate two different patterns of interface hydration that can be described as 'wet' and 'dry' (Janin, 1999). The antigen–antibody interface (Fig. 3) is wet, that is it is hydrated all through its surface. The Gα/G$\beta\gamma$ transducin interface (Fig. 4) is split into two distinct patches, both dry: many water molecules line their edges, but few if any penetrate inside.

Although the interface residues of a non-obligate complex are surface residues in the free components, there are some noticeable features in their amino acid composition. Relative to the accessible protein surface, the interfaces are depleted in Glu, Asp and Lys, and enriched in Met, Tyr and Trp. The propensities, shown in Fig. 7*c* together with those of protein–nucleic acid complexes discussed in Section 4.5, are generally weak, but they become more apparent once the interfaces are dissected into a core and a rim as in Fig. 7*a*. The rim, made of residues in which none of the interface atoms are fully buried, has a composition close to the protein accessible surface. The core comprises all the buried interface atoms, and 55% of all interface residues on average. It is enriched in aromatic residues, and to a lesser extent, in aliphatic residues, but not in Val, Ala and Pro. Val is indifferent, whereas Ala and Pro have a negative propensity for the interface core. Albeit aliphatic, the side chains of these residues are more constrained than in Ile, Leu, Met or the aromatic residues, and it may be that the flexibility of the longer side chains makes them more amenable to the needs of the interface formation. Arginine, far from being disallowed like the other charged residues, contributes 10% of the BSA in both the core and the rim, and also 10% of the ASA (Chakrabarti & Janin, 2002).

4.1.3 The core/rim model of protein–protein recognition

Interface residues make inter-molecular interactions. They are structurally and possibly functionally important, and therefore subject to a selection pressure that should be detected in homologous sequences. Several methods for predicting protein–protein contacts rely on identifying patches of surface residues conserved in evolution (Lichtarge *et al*. 1996; Armon *et al*. 2001; Valdar & Thornton, 2001; Lichtarge & Sowa, 2002; Ma *et al*. 2003). The conservation signal is usually weak, but like the amino acid composition signal, it can be enhanced by splitting the interfaces into core and rim regions. Guharoy & Chakrabarti (2005) estimated the effect of

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

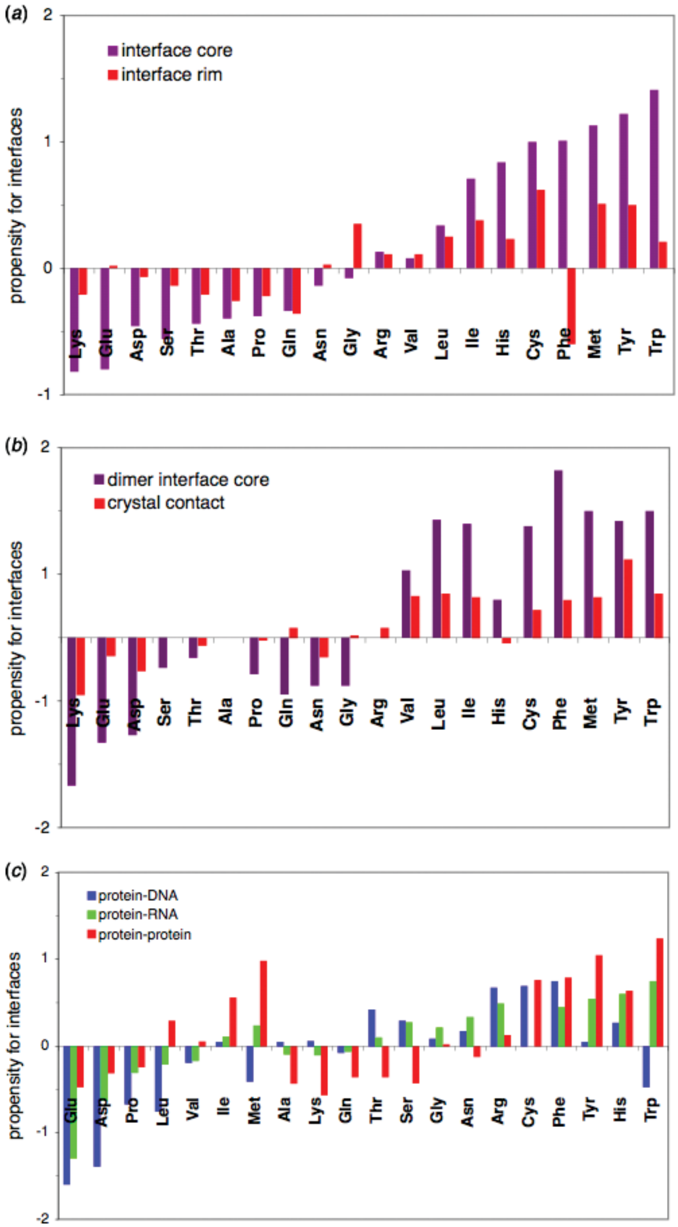


Fig. 7. Residue propensities for interfaces *vs.* the protein surface. The propensities [Eq. (6)] are derived from the relative contributions of the 20 amino acid types to the BSA of the interfaces and the ASA of the proteins in the same set. (*a*) Core and rim of the interfaces of protein–protein complexes. Data from Chakrabarti & Janin (2002). (*b*) The core of the homodimers interfaces is compared to crystal packing interfaces. Data from Bahadur *et al.* (2004). (*c*) The interfaces of protein–DNA (Nadassy *et al.* 1999) and protein–RNA complexes (Bahadur *et al.* 2008) are compared to those of protein–protein complexes (Chakrabarti & Janin, 2002).

evolutionary selection by measuring Shannon entropies in sets of aligned sequences. In protein–protein complexes, the mean value of the normalized entropy of the interface residues was $s=0{\cdot}65$ in the core and $s=0{\cdot}80$ in the rim (Table 2). Figure 8 shows how the core and the rim of

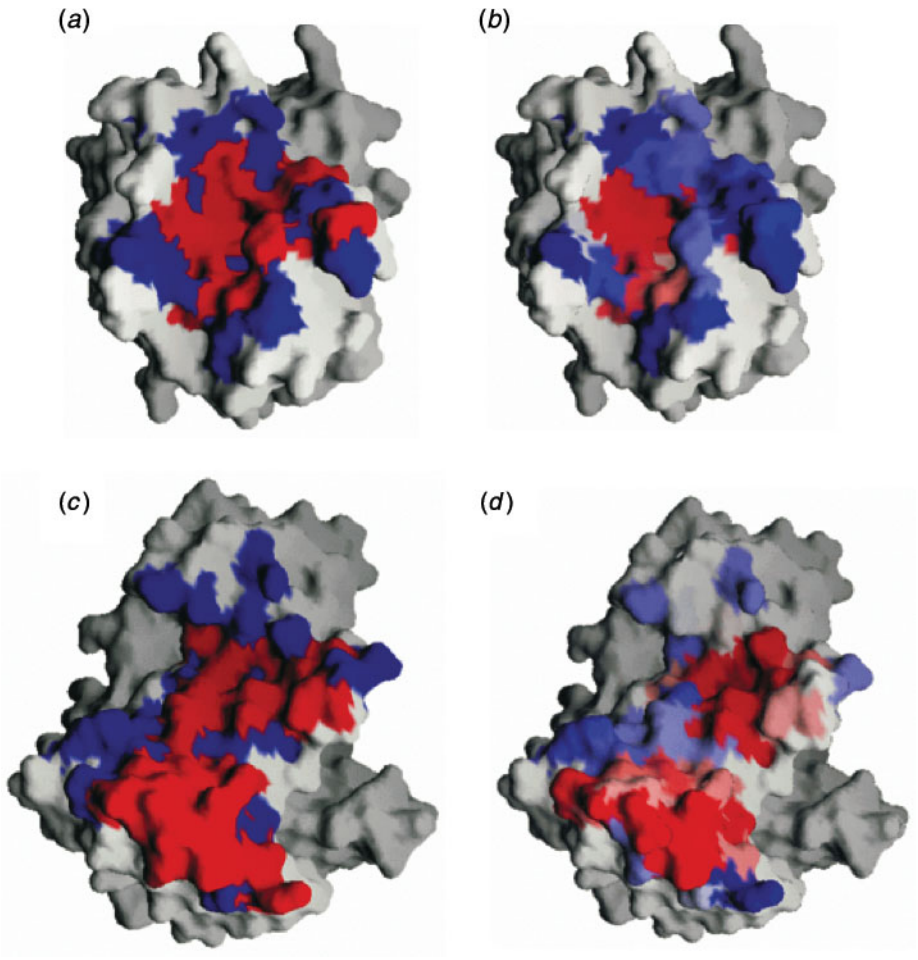


Fig. 8. The core/rim model of protein–protein interfaces. Top row: The protease component of the elastase–ovomucoid inhibitor complex (1ppf; Bode *et al.* 1986). Bottom row: The subunit of the malate dehydrogenase homodimer (1bmd; Kelly *et al.* 1993). (*a*, *c*) The core (red) comprises the residues that contain interface atoms with zero ASA, the rim (blue), the residues where all the interface atoms are accessible (Chakrabarti & Janin, 2002). (*b*, *d*) Shannon entropies of interface residues in aligned sequences [Eq. (7)]; red stands for low entropies (maximum conservation); blue denotes high entropies (maximum divergence). Figure made with GRASP (Nicholls *et al.* 1991), courtesy of M. Guharoy (Calcutta, India).

an interface are distributed on the surface of two proteins, and compares that distribution with the sequence entropy. The proteins are the inhibitor component of an enzyme–inhibitor complex, and the subunit of a homodimer.

The picture of the interface conservation derived from the Shannon entropy is coherent with the results of site-directed mutagenesis. Clackson & Wells (1995) noted that, in the HGH/HGH receptor system, the residues of the receptor that show large changes of affinity upon mutation to Ala are grouped at the center of the interface. In general, hotspot residues tend to cluster and be sheltered from the solvent (Bogan & Thorn, 1998; Halperin *et al.* 2004). Guharoy & Chakrabarti (2005) observed that in 14 complexes, there is a correlation between the contribution of the interface residues to the BSA and $\Delta\Delta G_{\mathrm{d}}$, the change in the free energy of dissociation of the complex upon their mutation to Ala. The correlation is significant for the interface core, but not

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

the rim, and it yields a slope of 26–38 cal mol^{-1} $\mathrm{\AA}^{-2}$, consistent with the estimates of Chothia (1974) and Vajda *et al.* (1995) based on the solubility of hydrocarbons. Sundberg *et al.* (2000) and Li *et al.* (2005) introduced a series of different size side chains at two sites of the interface of antibody–lysozyme complexes: a site at the center of the interface, the other at its periphery. In both series, $\Delta\Delta G_\mathrm{d}$ is proportional to the change in the BSA, but the slope is twice as large at the central site than at the peripheral site (46 *vs.* 21 cal mol^{-1} $\mathrm{\AA}^{-2}$).

Other features of the interfaces support the core/rim distinction, for instance the 'residue depth' of Chakravarty & Varadarajan (1999), who find it to be correlated to $\Delta\Delta G_\mathrm{d}$ values, or the distribution of the conserved 'dry residues' of Mihalek *et al.* (2007). Taken together, the site-directed mutagenesis and the sequence conservation data suggest that, in a complex, protein–protein recognition exerts most of its selection pressure on the interface core. The data fit an 'O ring' model in which the interface has a core of buried atoms and (partly) buried residues, surrounded by a rim of residues whose atoms remain solvent accessible (Bogan & Thorn, 1998; Lo Conte *et al.* 1999). The model applies as such to a single patch standard-size interface, for instance the antigen–antibody interface of Fig. 3, which is the one that was mutated by Sundberg *et al.* (2000), or the enzyme–inhibitor interface of Fig. 8*a*. In a multi-patch interface like that of transducin (Fig. 4), or the HGH/HGH receptor complex where the receptor is a homodimer and bears one recognition patch per subunit, each patch makes its own O ring.

To conclude, we wish to stress that the core/rim model, and the mutant data themselves, should not be interpreted to say that the rim plays no part in the interaction. The model and the data only state that the rim residues contribute less to the free energy of dissociation than the core residues, or that their contribution does not depend heavily on the nature of their side chain.

4.2 Oligomeric proteins

4.2.1 Interface size and stability in homodimers

Subunit interfaces have been analyzed in several sets of oligomeric proteins that comprise tetramers, hexamers and larger assemblies (Argos, 1988; Janin *et al.* 1988; Jones & Thornton, 1995, 2000; Ponstingl *et al.* 2005), but most of the available data concern homodimers, and we will limit our analysis to that category. Table 2 lists the mean values of the interface parameters in three separate sets of homodimers. That of Bahadur *et al.* (2003) was manually assembled and comprises 122 proteins. A second set more than twice that size, and a third set of 19 'weak' dimers were recently compiled by our group (Dey *et al.* unpublished data) with the help of an early version of the PiQSi database (Lévy, 2007); they are non-redundant at the 35% sequence identity level, and the QS of the proteins in them was confirmed by checking the literature. The weak dimers are homodimers reported to be in equilibrium with the monomers in the publications cited in PiQSi. Some of the homodimers in the first two sets may also be weak in that sense, but the scanned publications did not say. Noreen & Thornton (2003a, b) also assembled a set of weak dimers, but due to different criteria of choice, the present list shares only two entries with theirs.

The most obvious feature that distinguishes the homodimer interfaces in Table 2 is their large BSA, which on average is twice that of the non-obligate complexes. As the number of interface atoms and interface residues increases linearly with the BSA, the factor of 2 also applies to atoms and residues. However, the average BSA is of little significance given the very large standard deviation. Figure 9 shows that individual values spread from less than 1000 $\mathrm{\AA}^2$ to more than 6000 $\mathrm{\AA}^2$; the largest in the set is 14 300 $\mathrm{\AA}^2$. The BSA distribution is essentially the same in the Bahadur set and the larger new set, but it is very different in the weak dimer set. The mean BSA

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

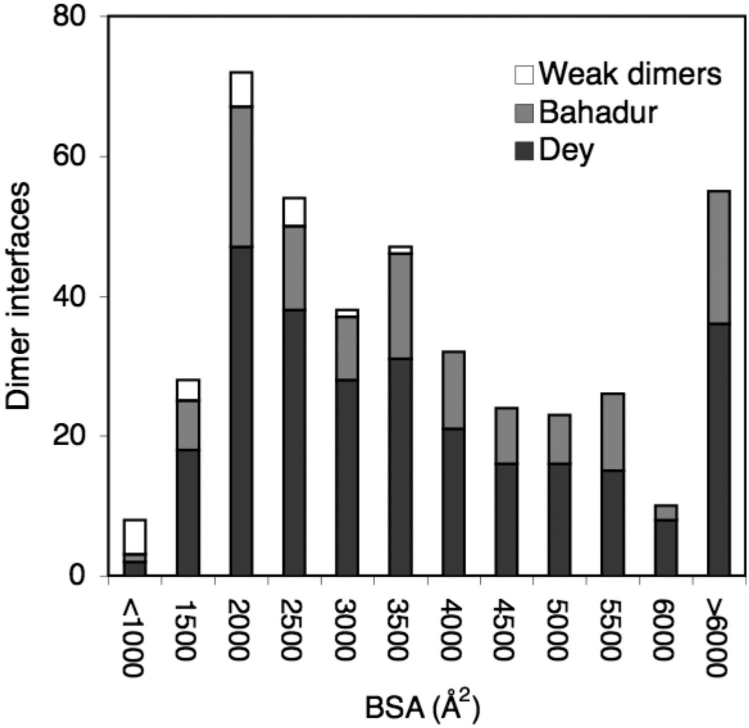


Fig. 9. Interface size in homodimers. Histogram of the BSA in the three sets of homodimers reported in Table 2; the weak dimers are in equilibrium with the monomers.

of the weak dimers is 1620 Å^2, consistent with the data of Noreen & Thornton (2003b), who report an average of 740 Å^2 for the buried surface area per subunit (that is, BSA/2) in their set of homodimers in equilibrium with monomers. In Fig. 9, most of the interfaces with BSA <1000 Å^2 belong to the weak dimers. The smallest, with a BSA near 750 Å^2, occur in cytochrome *c*6 (1c6o) and ferredoxin (1sj1), two proteins that may not be actual homodimers in the cell. The functional state of cytochrome *c*6 is likely to be a heterodimer (Schnackenberg *et al.* 1999); that of ferredoxin is unknown.

4.2.2 Chemical and amino acid composition

Non-polar (carbon-containing) groups contribute $f_{\mathrm{np}}=65\%$ of the BSA of homodimer interfaces, significantly more than to the accessible protein surface ($f_{\mathrm{np}}=57\%$), or to the interfaces in complexes; the weak dimer interfaces ($f_{\mathrm{np}}=62\%$) are in between. The hydrophobic character of the homodimer interfaces is associated in Table 2 with a lesser contribution of the neutral polar groups, but not of the charged groups, to the BSA. Nevertheless, homodimer interfaces do not, in general, form large hydrophobic patches (Larsen *et al.* 1998), and they contain many H bonds and water molecules. Table 2 reports an average of 18–19 H bonds and 44 water molecules per interface, which makes the surface density of the H bonds and the water molecules in homodimer interfaces similar to that of the complexes. The data also suggest that the H bond density is less in the weak dimers, but the statistics are poor because of the small size of that set.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

The amino acid composition of the homodimer interfaces has often been analyzed (Jones & Thornton, 1995, 1996; Tsai *et al.* 1997; Bahadur *et al.* 2003, 2004). It follows the same trends as the atomic composition: the interfaces are enriched in aliphatic and aromatic residues, on average by a factor of about 2, and depleted by the same factor in charged residues other than Arg. When the relative contributions of the residue types to the BSA and the ASA are expressed as propensities, the observation made above for the complexes remains valid: the propensities are much more marked for the residues of the interface core, which contain the buried interface atoms, than the interface rim. Figure 7 *b* reports the propensities to be at the interface core in the Bahadur set. A comparison with Fig. 7 *a* shows that they generally resemble those in complexes, but the aliphatic residues Ala, Val and Leu are more abundant. Leu and Arg are the largest contributors to the cores of homodimer interfaces, each with 10–11% of the BSA. In the complexes, Tyr and Arg both contribute 10% of the core BSA; Leu is third, but with only 6% (Chakrabarti & Janin, 2002; Bahadur *et al.* 2003).

The propensity of an atom or a residue to be at an interface *vs.* the protein surface may be viewed as one-body statistical potential. The R_p score of Bahadur *et al.* (2004), derived simply by summing the propensities over all the residues of an interface, has a positive and large mean value in homodimers (Table 2). It is also positive, but much lower, in the complexes and the weak dimers, in line with their relative stabilities. Two-body potentials can be derived statistically by counting pairwise contacts across the interfaces (Moont *et al.* 1999; Glaser *et al.* 2001; Ofran & Rost, 2003; Saha *et al.* 2005). When residue–residue contacts are analyzed in different types of protein–protein interfaces, the homodimers show a peculiar feature: pairs containing identical residues are much more frequent than expected on a random basis. This results from contacts made by twofold-related residues along the symmetry axis, which make up 12% of the BSA. After correcting for the relative abundance of each residue type, Leu comes to be by far the largest contributor to the pairwise contacts (13%), followed by Phe, Val and Arg (Saha *et al.* 2005). The leucine zippers are a well-known example of how Leu/Leu contacts may contribute to the stability of a homodimer.

4.2.3 Atomic packing and sequence conservation

The oligomeric proteins are generally known to assemble while the subunits fold, and their interfaces can be expected to be close-packed like the subunits' interior. Ponstingl *et al.* (2005) observe that the Voronoi volumes of the interface atoms of oligomeric proteins are within a few percent of the reference volumes of Tsai *et al.* (1999), but this applies only to the buried atoms, which are a minority (f_{bu} = 35–36% in Table 2). Buried atoms are even fewer at the interfaces of weak dimers (f_{bu} = 28%). Table 2 lists two other parameters related to the atomic packing: the L_D packing index of Bahadur *et al.* (2004) and the S_c shape complementarity score of Lawrence & Colman (1993). Both take very similar mean values in homodimers and in complexes, which supports the idea that all these interfaces are close-packed. On the other hand, the weak dimer interfaces have low values of f_{bu} and of L_D, and they may be poorly packed.

Guharoy & Chakrabarti (2005) analyzed the conservation of the interface residues in homodimers, and obtained a result very similar to what they had seen in complexes (Table 2): the mean Shannon entropy of interface residues is significantly less than the average in the whole polypeptide chain, and it is lower for the residues of the interface core than the rim. In that study, the mean number of residues in the core was the same in both types of interfaces, and the core represented a majority of the residues. This suggests that the core/rim model of interfaces is also

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

valid for homodimers, but the correlation with hotspots is difficult to establish in their cases. Most are permanent assemblies much less amenable to a quantitative study than a non-obligate complex. The monomers are in general not observed, and K_d values cannot be measured. The measurement can, in principle, be done on the weak dimers, but mutagenesis data are comparatively scarce on those.

4.3 Non-specific interactions in crystals

4.3.1 Size and composition of crystal packing interfaces

Protein crystals are held by the same forces that stabilize other macromolecular assemblies, and one may compare the interfaces between pairs of molecules in crystals to those of the complexes and oligomeric proteins (Janin & Rodier, 1995; Carugo & Argos, 1997; Dasgupta *et al.* 1997). A feature then strikes the eye: their small size. In the set of 1320 pairwise interfaces assembled by Janin & Rodier (1995), the mean BSA is 570 Å^2. Crystal packing nevertheless buries much protein surface, because each molecule has 8–10 neighbors, and it makes pairwise interfaces with all. The average crystal packing interface represents only one-third of that of a complex, and one-sixth of that of a homodimer. The great majority buries less than 800 Å^2 and implicates fewer than 13 residues per protein molecule. It is often assumed that such interfaces cannot be biologically meaningful, but there are some larger ones in crystals. The BSA values approximate an extreme value distribution, in which large values are much more frequent than in a Gaussian distribution. Moreover, the interfaces with twofold symmetry tend to be larger than those created by other types of crystal symmetry operations (Janin, 1997). About 10% of the crystal packing interfaces are similar in size to those of the complexes, and a majority of those have twofold symmetry. They form 'crystal dimers', which are artifacts of crystallization and do not exist in solution, yet they may be difficult to distinguish from biological homodimers on the basis of the crystal structure alone.

Bahadur *et al.* (2004) selected 188 large crystal packing interfaces by keeping only those with a BSA $>$800 Å^2 in the set of Janin & Rodier (1995). They have an average BSA of 1510 Å^2, comparable to standard-size interfaces in complexes, and they comprise about the same average number of atoms and of residues (Table 2). Their chemical composition is also similar to that of the biologically significant interfaces, except for a slight excess of charged groups. Their amino acid composition is reported in Fig. 7*b* in the form of propensities that are all small ($<0{\cdot}5$) in absolute value, but follow the same pattern as homodimers and complexes: there is a mild depletion in Lys, and a mild excess of aliphatic and aromatic residues. The R_p score, which is the sum of the propensities of individual interface residues to be at a homodimer interface and has a positive mean value in complexes, is negative on average at crystal contacts (Table 2).

Crystal packing interfaces have a comparatively low average density of H bonds: 1 per 280 Å^2 of BSA, instead of 1 per $\approx$200 Å^2 in complexes and homodimers (Table 2). This may be related to their low fraction of buried atoms: $f_{bu}=21\%$ *vs.* 34–36%; and their high level of hydration: 15 waters per 1000 Å^2 of interface area *vs.* 10–11 in homodimers and complexes. Their topology is another cause for these differences. A standard-size interface in a complex is in general single-patch and compact, and a crystal packing interface of the same size is almost always fragmented. This can be seen in Fig. 10, which compares the interfaces of a biological homodimer (κ-bungarotoxin, PDB entry 1kba) and a crystal dimer (pokeweed antiviral protein, PDB entry 1qci). The two are of similar size, but the second is poorly packed; it buries very few

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

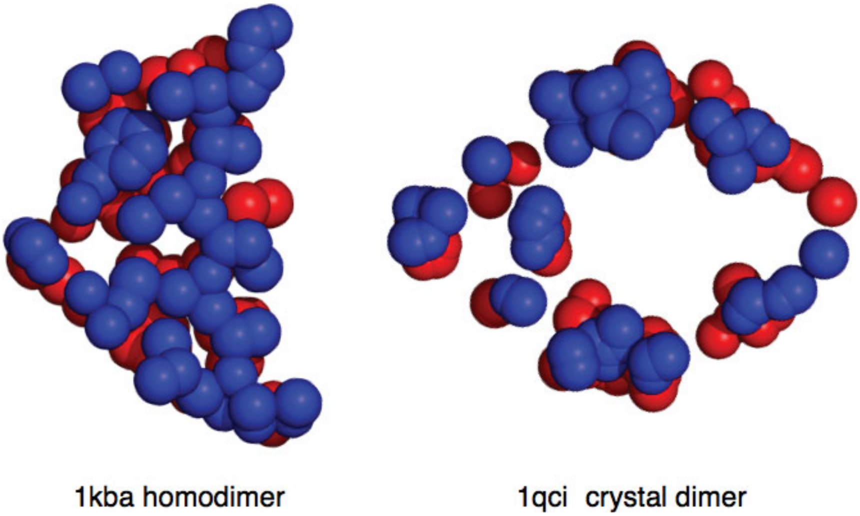


Fig. 10. The atomic packing of a specific and a non-specific interface. The κ-bungarotoxin homodimer (1kba) and the pokeweed antiviral protein crystal dimer (1qci) form interfaces of a similar size, with a BSA of almost 1000 Å^2, but only the first is biologically significant. Each subunit of either dimer contains about 50 interface atoms, drawn here as spheres in the plane of the twofold axis. Their packing is very different, as are the values of the fraction of buried atoms ($f_{\mathrm{bu}}=28\%$ *vs.* 7%) and the L_D index (31 *vs.* 13).

atoms and its L_D index is low. In Table 2, the fragmentation of crystal packing interfaces shows in the low mean value of L_D and in the number of polypeptide segments, which is greater on average at a crystal packing interface than at an interface of the same size in a complex; homodimer interfaces implicate an even smaller number of segments in proportion of their size.

4.3.2 Biological assemblies *vs.* crystal artifacts

These differences may be used to identify crystal and biological dimers in PDB entries. The interface size alone correctly distinguishes homodimers from monomers in 85% of the cases (Ponstingl *et al.* 2000). Additional information can be drawn from the literature, residue conservation or physical–chemical properties of the interfaces as discussed above (Section 2.3). In the sample of Bahadur *et al.* (2004), where all the crystal packing interfaces have a BSA >800 Å^2, size is not a powerful discriminator, but it can be combined with three parameters that tend to take higher values in biological homodimers than crystal dimers: the fractions of non-polar groups (f_{np}) and buried atoms (f_{bu}), and the number of H bonds (n_{HB}). Thus, an interface has an 88% probability of belonging to a biological homodimer if it satisfies one of the two conditions:

$$f_{\mathrm{np}}\cdot \mathrm{BSA} > 1000\ \text{Å}^2 \text{ and } f_{\mathrm{bu}}>24\%$$

$$f_{\mathrm{np}}>61\% \text{ and } n_{\mathrm{HB}}>8$$

and it has the same probability to be due to the crystal packing if neither condition is met (Janin *et al.* 2007).

The molecular contacts in crystals are unspecific and not biologically meaningful, with a few interesting exceptions (Janin, 1997). Unlike specific interfaces, they should exert no evolutionary

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

pressure on the regions of the protein surface implicated, an assumption supported by a comparison of aligned sequences. Table 2 reports the normalized Shannon entropy of the residues in the large crystal packing interfaces. Its mean value is $s \approx 1$, meaning that they evolve at the same rate as the average residue in the polypeptide chain to which they belong. Moreover, the core and the rim have nearly identical s, the core and rim residues being defined here by the buried atoms as in the other types of interfaces even though many crystal packing interfaces do not have a proper core, because they bury few atoms and are highly fragmented. In addition, proteins often come in several crystal forms that use different contacts. Pancreatic ribonuclease, for instance, crystallizes in at least six forms. Several contain the same large interface that forms a dimer in solution under the conditions that lead to crystallization, but the other crystal packing contacts essentially implicate the whole protein surface (Crosio *et al.* 1992).

4.4 Icosahedral virus capsids

4.4.1 Symmetry

Icosahedral virus capsids are multi-subunit assemblies, a category that constitutes most of the cellular machines, but is poorly represented in the PDB at present (Dutta & Berman, 2005). The capsids encapsulate and protect the viral genome, and they are frequently implicated in the recognition and infection of target cells. They are very large objects with molecular weights of millions, and a symmetry that greatly helps in their study, so that atomic structures could be solved three decades ago for tomato bushy stunt virus (Harrison *et al.* 1978) and satellite tobacco necrosis virus (Liljas *et al.* 1982). Here, we can compare the subunit interfaces in capsids to those found in binary assemblies in terms of their structural and physical chemical properties, and discuss the process of capsid self-assembly.

The virus capsids were considered spherical until Crick & Watson (1956) proposed that they have the symmetry of the cubic I (icosahedral) point group. This was demonstrated by the electron microscopy studies of Caspar & Klug (1962), who introduced the rule of quasi equivalence and the lattice triangulation number T. Quasi equivalence allows the icosahedral asymmetric unit (IAU) to contain T subunits, and the whole capsid, $60T$ subunits instead of the 60 implied by the point group. The subunits in the IAU are related by inexact or 'quasi' symmetries. They have identical sequences, but different conformations, and they make different contacts with their neighbors (Rossmann & Johnson, 1989). X-ray structures are available for a number of capsids with the $T=1$ lattice exemplified by satellite tobacco necrosis virus, the $T=3$ lattice exemplified by tomato bushy stunt virus, and also the *pT3* (pseudo $T=3$) lattice exemplified by rhinovirus (Rossmann *et al.* 1985). The *pT3* capsids resemble the $T=3$ capsids, but the subunits in their IAU have different sequences. The PDB also reports the X-ray structures of a few capsids with larger T numbers, or with lattices that do not follow the rule of quasi equivalence, for instance the very large and complex capsids of Bluetongue virus (Grimes *et al.* 1998) and phage PhiX174 (Dokland *et al.* 1999). The VIPER database of the Scripps Research Institute (Shepherd *et al.* 2006) offers a particularly convenient access to some 250 virus structures obtained by X-ray and electron microscopy.

4.4.2 Subunit interfaces in capsids

Bahadur *et al.* (2007) have assembled a non-redundant dataset of 49 viral capsids that includes 11 with lattice $T=1$, 17 with lattice $T=3$, 10 with lattice *pT3* and 11 others. In each capsid, a unique

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

set of interfaces between pairs of polypeptide chains can be identified. Icosahedral symmetries repeat these unique interfaces 60 times, except those with twofold (I_2) symmetry, which are repeated 30 times. There is an average of 16 unique interfaces per capsid; one-third occurs within the IAU, another third has I_2, I_3 or I_5 symmetry, the remainder, a quasi-symmetry.

The average pairwise interface in that set has a BSA of 1750 Å^2, similar to transient complexes, but the range of sizes is large. Bahadur *et al.* (2007) split the pairwise interfaces into three size categories. The small interfaces (BSA <800 Å^2), which make up 40% of the sample, contribute only 7% of the BSA. The remainder belongs to medium-size (800–2000 Å^2) and large interfaces (more than 2000 Å^2). The $T=1$ capsids contain in general three unique interfaces with a BSA >800 Å^2, one of each of the I_2, I_3 and I_5 symmetry types; $T=3$ capsids contain 5–9, including 3 between the chains in the IAU; *pT3* capsids contain 6–16.

In a virus capsid, each subunit has many neighbors and buries a large fraction of its surface in contacts with them. The average number of neighbors is 7 in $T=1$ and $T=3$ capsids. In these capsids, the subunit contacts implicate about 60% of all residues and bury about 45% of the subunit ASA. In *pT3* capsids, the number of neighbors increases to 12; the fraction of the residues that are at interfaces, to 73%; and the fraction of the ASA that is lost, to 60%. The complexity of the capsid assembly and the multiplicity of interfaces of all sizes are illustrated in Fig. 11 for rhinovirus (PDB entry 1 aym; Hadfield *et al.* 1997). Its *pT3* capsid contains four different polypeptide chains, one of which (chain A in gray) forms homopentamers about the icosahedral fivefold axes, whereas chains B (blue) and C (cyan) form heterohexamers about the threefold axes. Chain A covers much of the outer surface of the capsid, yet the bottom part of Fig. 11 shows that in two opposite orientations, most of its surface is involved in one or several of the 13 different interfaces it makes with its neighbors.

An important feature of the capsid assembly cannot occur in binary complexes and homodimers: the pairwise interfaces between subunits overlap, and many atoms or residues are part of more than one. On average, 15% of the interface atoms and 24% of the interface residues are in contact with two neighboring subunits, 2% of the atoms and 5% of the residues with three, and there are extreme cases where a residue is in contact with seven subunits (Bahadur *et al.* 2007).

4.4.3 Composition and topology

Capsid interfaces resemble homodimer interfaces in their non-polar character ($f_{\mathrm{np}}=63\%$), irrespective of their size (Bahadur *et al.* 2007). The fraction of buried atoms (f_{bu}) is only 29% in the pairwise interfaces, but it increases to 36% in the whole assembly. The difference comes from atoms in contact with more than one neighboring subunit; they can have a non-zero ASA in each of the pairwise interfaces, and a zero ASA in the capsid. The large capsid interfaces have f_{bu} and L_D packing indexes similar to those of homodimers, which implies that they are well packed. In contrast, the interfaces with BSA <2000 Å^2 bury very few atoms ($f_{\mathrm{bu}}=11$–23%), and their L_D index is low. Like crystal packing contacts, they may be loosely packed.

The interface core, which contains the buried atoms, represents half of the residues in pairwise interfaces, and two-thirds in the whole capsid. The high proportion of core residues affects the amino acid composition of the capsid interfaces. As in homodimers and complexes, the core is enriched in aliphatic residues and depleted in Asp, Glu and Lys, but not Arg. On the other hand, the capsid interfaces contain fewer aromatic and more neutral polar residues: Ser, Thr, Asn, Gln and Pro contribute 32% of the BSA in capsids *vs.* only 21% in homodimers.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

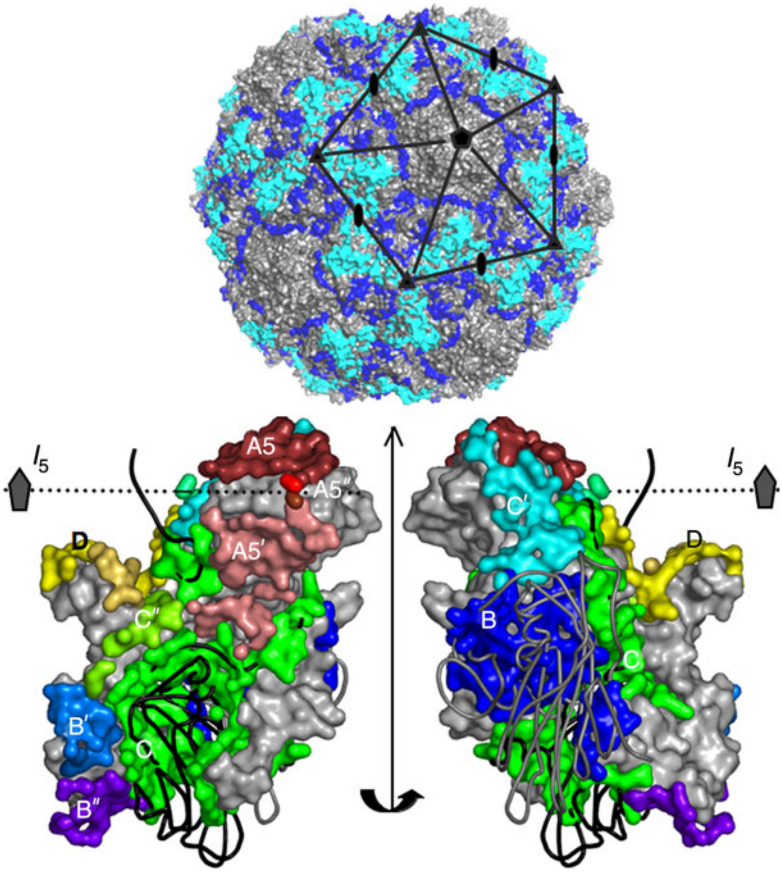


Fig. 11. Capsid assembly and subunit interfaces in rhinovirus. Rhinovirus (1aym; Hadfield *et al.* 1997) has a *pT3* capsid with four polypeptide chains in the IAU. Top: The capsid and its icosahedral lattice; chain A in gray forms pentamers about the fivefold axes; chain B in blue and chain C in cyan form heterohexamers about the threefold axes. Bottom: The molecular surface of chain A is drawn in two orientations 180° apart in the plane of the I_5 axis (dashed line), and colored according the subunit contacts. Chain B is drawn as a gray tube, chain C as a black tube. In the capsid, chain A has 13 neighbors with which it makes five large, four medium-size and four small interfaces. The large interfaces are with chain B (blue surface) and chain C (green), two A chains in the A pentamer (pink and brown) and chain C′, threefold related to C (cyan). The medium size interfaces involve the small chain D (yellow) and symmetry-related chains B′, B″ and C″. Chain A5″, related to A by a 144° rotation about I_5, makes a small interface (red) that implicates only one residue of A. Figure adapted from Bahadur *et al.* (2007).

Capsid interfaces contain an average of 7 H bonds. The surface density of 1 H bond per 250 Å^2 of BSA is less than for the homodimers in Table 2, but the difference is probably an artifact of the comparatively low resolution of many of the viral X-ray structures: the interface H bond density is 1 per 200 Å^2 in the 17 capsids structures with 2·8-Å or better resolution.

The topology of capsid interfaces was analyzed with the clustering method of Chakrabarti & Janin (2002). Most of the small interfaces are single-patch, medium size interfaces form one to three pairs of patches and large interfaces three or more. Interfaces of equivalent size in complexes have a similar number of patches. When the interface residues are distributed into segments of the polypeptide chain following Jones & Thornton (1997) and Pal *et al.* (2007), the mean number of segments is 3·9 per chain, and that of interface residues per segment is 6·4, again like in the complexes in Table 2. However, the averaging hides a great disparity between the small interfaces, often fragmented into very short segments, and larger ones that tend to contain a small number of long segments. In $T=3$ and *pT3* capsids, some interfaces contain segments of up to 47 residues located at the N or C terminus of the polypeptide chain. These tail segments adopt extended conformations and are heavily involved in subunit contacts. In $T=3$ capsids, the tails are preferentially involved in the twofold and quasi-sixfold interfaces; in *pT3* capsids, they contribute primarily to the interfaces in the IAU (Bahadur *et al.* 2007).

4.4.4 Residue conservation

The subunit contacts hold the capsid together and play a major role in the process by which it assembles itself. They should therefore impose stringent evolutionary constraints on the protein sequence. Bahadur & Janin (2008) used the Shannon entropy to estimate residue conservation in sets of aligned capsid protein sequences. The mean value of s normalized to the average value in each polypeptide chain, is close to 1 for the interface residues, which reflects the fact that they constitute two-thirds of the polypeptide chain. Nevertheless, there are very significant differences within a chain. As in other systems, the residues of the protein interior are much better conserved than surface residues ($s=0{\cdot}7$ *vs.* 1·6), and the residues of the interface core, much better conserved than the rim ($s=0{\cdot}8$ *vs.* 1·2). Moreover, the capsids contain residues involved in several interfaces, with no equivalent in homodimers or binary complexes. They are abundant (34% of all interface residues), and their mean entropy $s=0{\cdot}8$ shows that they are better conserved than the residues involved in only one interface.

The sequence conservation needs not be homogeneous within a given interface. An example is the interface between chains A and C of the *pT3* rhinovirus (the green surface in Fig. 11). It is well conserved as a whole ($s=0{\cdot}9$), but chain A has a highly divergent C tail with $s=2{\cdot}0$ and a highly conserved N tail with $s=0{\cdot}6$, and both tails contribute to the interface. The C tails are more divergent than the N tails in most *pT3* capsids, but in general, the tails have the same rate of evolution as the parts of the polypeptide chains that form the globular core of the subunits (Bahadur & Janin, 2008).

4.4.5 A plausible mechanism for capsid assembly

The self-assembly of a virus capsid comprises the folding of its subunits, their association and maturation steps that often involve conformation changes and covalent modifications. This elaborate process may implicate nucleic acids, accessory viral proteins and host chaperones, in addition to the protein subunits (Steven *et al.* 1997, 2005; Liljas, 1999; Caspar, 1980). Nonetheless, it can be amazingly fast: bacteriophage T4 goes through a complete cycle of infection, replication and lysis in 15 min; the capsid assembles in minutes in spite of the large number of polypeptide chains that constitute it (Leiman *et al.* 2003). This implies that capsid self-assembly must go through a series of low-order steps and intermediate species. These species, called capsomers, should resemble small oligomeric proteins.

Prevelige *et al.* (1993) have analyzed the *in vitro* self-assembly of bacteriophage P22, and Xie & Hendrix (1995), that of bacteriophage HK97. The $T=7$ lattice of these two capsids can be viewed as comprising pentamers with I_5 symmetry and hexamers with Q_6 quasi-symmetry, in a 1:5 ratio. In solution, the HK97 subunits form pentamers and hexamers that interconvert slowly and reassemble most efficiently when they are in the same 1:5 ratio as in the capsid. All models of

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

the capsid self-assembly assume that the capsomers keep their structure during the process, and thus, they can be identified in the capsid itself (Johnson & Speir, 1997; Reddy *et al.* 1998; Dokland, 2000; Zlotnick, 2005). Bahadur *et al.* (2007) postulate in addition that the capsomers contain the largest interfaces, and that capsomer–capsomer association makes use of the medium-size interfaces in priority. The many small interfaces that are seen in the capsids may contribute to the stability of the final product, but they are unlikely to play a role in its formation.

In the case of HK97, this postulate suffices to determine an assembly pathway compatible with the data of Xie & Hendrix (1995). The largest interfaces are the I_5 and Q_6 interfaces, and they build the hexamers and pentamers seen in solution. The next largest are of medium size; they occur between hexamers or between hexamers and pentamers in the capsid, and in solution, they should allow hexamers to associate in pairs or with a pentamer (Bahadur *et al.* 2007). Recently, Stockley *et al.* (2007) have analyzed the assembly of bacteriophage MS2 by electrospray ionization-mass spectrometry. They find that the $T=3$ capsid of MS2 dissociates into symmetric dimers, some of which become asymmetric upon addition of an RNA stem–loop fragment. Symmetric and asymmetric dimers then associate into dimer hexamers, whereas no pentamer is formed. These observations are compatible with the model of Bahadur *et al.* (2007): MS2 has large I_2 and Q_2 interfaces that build a symmetric and an asymmetric dimer, respectively. The next largest are the Q_3 interfaces building the dimer hexamers; they are of medium size, and larger than the I_5 interface needed to build pentamers.

4.5 Protein–nucleic acid recognition

Nucleic acid recognition by proteins is a process of major interest to biology, actively studied by biophysicists, structural biologists and bioinformaticians. In this review, we consider it only in relation to protein–protein recognition, and use results of the transverse studies of protein–DNA complexes by Jones *et al.* (1999), Nadassy *et al.* (1999) and Sarai & Kono (2005), and of protein–RNA complexes by Jones *et al.* (2001), Treger & Westhof (2001), Ellis *et al.* (2007) and Bahadur *et al.* (2008). Each study relies on sets of PDB entries that comprise 26–81 complexes, similar in size to the sets of protein–protein interfaces discussed in previous sections. The DNA is double stranded in most of the complexes; the RNA is single stranded with few exceptions, but when its sequence and length allow, it folds into stem–loops and a variety of other structures that include double helical segments. The complexes are mostly non-obligate, forming only when the protein encounters the nucleic acid, but like the protein–protein complexes of Section 4.1, they cover a wide range of stability, lifetimes and functions.

The data in Table 3 show that protein–DNA and protein–RNA complexes tend to bury less molecular surface than homodimers, which are obligate assemblies, but more than the average protein–protein complex. Here, again, the presence of large interfaces is associated with conformation changes that affect the protein and the nucleic acid components of many of the complexes (Jones *et al.* 1999; Nadassy *et al.* 1999; Ellis & Jones, 2008). A peculiar feature of protein–RNA complexes is that more surface area is lost on the RNA than the protein side: 8% on average, which is like the distribution of the BSA at the interfaces of protease–inhibitor complexes, and unlike most other types of protein–protein interfaces (Bahadur *et al.* 2008). The asymmetry, attributable to a convex RNA surface fitting into a concave protein surface, is low in protein–DNA complexes, and most pronounced when the RNA is short and does not form secondary structure.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Table 3. *Properties of protein–nucleic acid interfaces*

Average value	Protein/RNA[a]	Protein/DNA[b]	Protein/protein[c]
Number of complexes	81	75	70
BSA (Å^2)			
Mean	2530	3100	1910
S.D.	(1210)	(1050)	(760)
Protein/nucleic acid	1210/1320	1540/1560	–
Number of amino acids/nucleotides	43/18	48/18	57
BSA (Å^2) per amino acid/nucleotide	28/75	33/72	34
Composition (protein/nucleic acid, BSA %)			
Non-polar	55/33	52/41	58
Neutral polar	21/41	24/16	28
Charged (negative)	4/26	2/43	5
Charged (positive)	20/0	23/0	9
f_{bu} (% buried atoms, protein/nucleic acid)	29/29	24/28	34
L_D (packing index, protein/nucleic acid)	37/43	39/46	42
H bonds			
n_{HB} (number per interface)	20	22	10
BSA per bond (Å^2)	125	145	190
Water molecules			
Number per interface	32	21	20
Number per 1000 Å^2	13	7	10
Bridging H bonds	11		6

[a] Data from Bahadur *et al.* (2008). Numbers in the left column are for the protein component; those in the right column are for the RNA component.

[b] Data from Nadassy *et al.* (1999). Numbers in the left column are for the protein component; those in the right column are for the DNA component.

[c] Data reported from Table 2 for comparison.

The BSA histograms of Fig. 12 show that the protein–nucleic acid interfaces have a wide range of size. Both DNA and RNA can form very large interfaces with proteins. The peak of the distribution is near 3000 Å^2 for DNA and 2000 Å^2 for RNA. Many of the proteins that bind DNA are homodimers or tandem repeats that bear two or more sites of interaction: 60% in the set of Nadassy *et al.* (1999). This is less common with RNA binding proteins and may explain the larger average BSA of the protein–DNA interfaces. Moreover, while very few protein–DNA interfaces bury less than 1500 Å^2, 10% of the protein–RNA interfaces have a BSA <1200 Å^2. The smallest, with a BSA near 900 Å^2, occur in crystals with a very short RNA component that may reproduce only part of the protein–RNA contact in the cell. Omitting those, the distributions in Fig. 12 suggest that there is a minimum size for a functional protein–nucleic acid interface, and that it has the same BSA as the smallest functional protein–protein interfaces. Thus, protein–nucleic and protein–protein recognition have a similar size rule, which extends from the BSA to the number of amino acid residues implicated in recognition, since the BSA per interface residue is nearly the same (Table 3). On the DNA and RNA side of the interfaces, the nucleotides engaged in secondary structure tend to contribute less to the BSA than those in extended segments, probably because they are less solvent accessible to start with.

Beyond these similarities, major differences between the interfaces of protein–DNA or RNA complexes on one hand, and protein–protein complexes on the other hand, reflect the different chemical nature of proteins and nucleic acids. The bases, the sugars and the phosphate groups

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

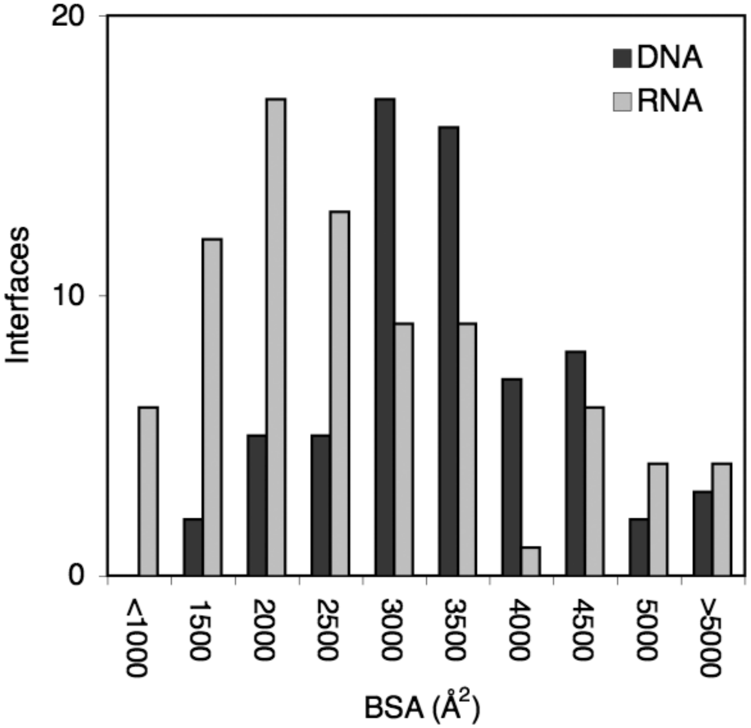


Fig. 12. Interface size in complexes with DNA and RNA. Histogram of the BSA in the sets of protein–DNA and protein–RNA complexes reported in Table 3. The BSA is from both the protein and the nucleic acid components.

contribute about one-third each to the ASA of the nucleic acids, but their respective contributions to the BSA are different in DNA and RNA. The phosphates contribute 43% of the DNA surface buried in contacts with proteins, and the deoxyribose, 29%. With RNA, the phosphate contribution is less (26%), and that of the ribose reaches 39% because of the heavy implication of the 2′-OH in interactions with protein groups (Treger & Westhof, 2001; Bahadur *et al.* 2008). The DNA or RNA side of the interfaces is highly polar and negatively charged, and in counterpart, the protein side is more polar than a protein–protein interface or than the solvent accessible surface: f_{np} is 52–55% instead of 57%. Moreover, it is positively charged over 20–23% of its BSA, which implies a peculiar amino acid composition and marked propensities. The propensities to be in contact with DNA are more marked than with RNA, and they differ from the propensities to be in contact with another protein, plotted for comparison in the same Fig. 7*c*. Asp and Glu are nearly excluded from contacts with both DNA and RNA; Arg is favored, whereas Lys is indifferent, but abundant. Aromatic residues are mildly favored at protein–RNA, but not protein–DNA interfaces; aliphatic residues are either indifferent (Val, Ile) or disfavored (Leu, Met, Pro).

Protein–nucleic acid interfaces bury an even lower fraction of their interface atoms than protein–protein interfaces do ($f_{\text{bu}}=24$–29% *vs.* 34–36%; Tables 2 and 3) and their L_D packing index is lower (37–39 *vs.* 42–45). This suggests that on average, the atomic packing is less tight than at protein–protein interfaces, yet the buried atoms on the protein side of protein–DNA interfaces have Voronoi volumes that are within 5% of the volumes in the protein interior (Nadassy *et al.* 1999). On the DNA side, the buried atoms have volumes greater by 0–8% than

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Table 4. *Euclidean distances between amino acid compositions*

Interface/surface	Protein surface	Protein interior
Protein surface		3·8
Homodimer interfaces[a]		
Core	3·9	1·5
Rim	2·0	
Interfaces in complexes[b]		
Core	3·4	2·7
Rim	1·2	
Crystal packing interfaces[c]	1·6	
Protein–DNA interfaces[d]	3·9	5·8
Protein–RNA interfaces[e]	3·4	

The Euclidean distance Δf between the percent composition f of the protein surface or interior and that, f', of an interface, is defined by $\Delta f^2 = 1/19\sum_{i=1-20}(f_i - f_i')^2$. The compositions are expressed as percent contributions to the ASA or the BSA.

Data from [a]Bahadur *et al.* (2003), [b]Chakrabarti and Janin (2002), [c]Bahadur *et al.* (2004), [d]Nadassy *et al.* (1999) and [e]Bahadur *et al.* (2008).

reference volumes derived from crystals of pure DNA (Nadassy *et al.* 2001). The packing quality expressed by the f_{bu} and L_D parameters is particularly poor in the complexes that involve transfer RNA, even though their interfaces are large (Bahadur *et al.* 2008).

The polar character of protein–nucleic acid interfaces correlates with their large number of H bonds and high residual hydration (Table 3). The surface density of the H bonds is significantly greater than in protein–protein complexes (1 per 125–145 Å^2 *vs.* 190 Å^2), in line with the contribution of polar atoms to the BSA of each system; and also that of the interface waters, but the values in Table 3 suffer from the unequal resolution of the X-ray structures and the inconsistent way they report solvent positions.

5. Conclusion: folding and recognition

Protein–protein recognition has much in common with protein folding. Both are self-assembly processes during which solvent is removed from the surface of polypeptide chains, whereas new van der Waals and polar interactions are formed between protein atoms. Features that stress that similarity are the atomic packing of the interfaces, their amino acid composition and their conservation in evolution. All three distinguish between the core and the rim of the interfaces. The packing density judged from the Voronoi volumes is the same for the atoms buried inside proteins and at the interfaces, where the buried atoms define the interface core. We have seen that the residue conservation in homologous sequences estimated by the Shannon entropy is a property of the interface core as opposed to the rim. In oligomeric proteins, the amino acid composition of the interface core resembles that of the protein interior, remote from that of the solvent accessible surface. This is presented in Table 4 in the form of Euclidean distances. The greatest distances are between the protein surface and either the interface core or the protein interior. The interface core of protein–protein complexes is equally distant in composition from the protein surface and the interior, in line with the fact that the residues concerned are surface residues in the free components, and interior residues in the complexes. The rim of both types of interfaces, and the crystal packing interfaces, which bury few atoms and do not have a

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

well-defined core, are all close in composition to the protein surface. The protein surface in contact with DNA and RNA has a very different composition from all the other surfaces, but that can be safely attributed to the different chemical nature of its partner.

Several other features do, however, distinguish protein–protein interfaces from the protein interior. One is the presence of immobilized water. Its abundance, which is certainly underestimated in Table 2, suggests that the dehydration of protein–protein interfaces is only 80–90% complete, whereas that of the protein interior is almost 100% (Hubbard & Argos, 1994). Another concerns arginine residues. Their highly polar side chain is essentially excluded from the protein interior, but not of protein–protein interfaces. Arginine is abundant in all types of protein–protein interfaces, even in their core. Some interface arginines are essential to the biological function, for instance the P1 arginine of trypsin inhibitors (Janin & Chothia, 1976), or the 'arginine finger' of Ras-GAP that activates the hydrolysis of GTP by the Ras protein (Scheffzek *et al.* 1997); others just contribute to stability. Birtalan *et al.* (2008) have recently analyzed the role of arginines and tyrosines at the antigen-combining site of antibodies, where they are the most abundant residue types. Residues of both types make stabilizing interactions, but arginine discriminates poorly between the cognate antigen and other proteins. This reminds us of the dual role of arginine in protein–DNA recognition: its side chain gives H bonds both to the phosphate backbone and to guanine bases in the major groove. The first H bonds are non-specific, yet important for stability; the second are major elements of the sequence specificity of many transcription factors.

The analogy between subunit assembly and protein folding is particularly significant when the assembly undergoes large conformation changes, that is in flexible recognition. We observe that flexible recognition often leads to the formation of a large interface, whereas rigid-body recognition yields a small or a standard-size interface. The changes that accompany association in that case resemble late stages of protein folding: preformed elements, secondary structure or domains, move one relative to the others. The movements affect the shape of the protein surface, which raises the question of how specificity can be achieved, and at what stage of the assembly it appears. Disorder-to-order transitions, which represent early steps in folding, are the rule in homodimers, where folding and assembly occur simultaneously. They are a common event in proteins that bind DNA, and also in protein–protein complexes that have a natively denatured component. In our sample, such transitions are observed only locally, for instance at the N terminus of Gα in transducin.

The viral capsids display some remarkable features that they probably share with other multicomponent assemblies, but not binary complexes or homodimers. In a capsid, the subunit contacts bury half or more of the protein surface; in complexes or homodimers, that fraction rarely exceeds 20%. The contacts implicate a majority of the protein sequence, and they affect its composition and its conservation in evolution much more than in a smaller assembly. Many interfaces of very different sizes coexist, and they frequently overlap, with one-third of the residues part of two or more interfaces; these residues are much better conserved than those involved in only one interface. The size distribution of the interfaces has led us to propose a simple model of the capsid assembly, in which the largest interfaces form first. This model, compatible with experimental data on two bacteriophages, may be extended to other multicomponent systems: recent data from mass spectrometry indicate that the assembly of some large oligomeric proteins also proceeds by forming the largest interfaces first (Lévy *et al.* 2008). The self-assembly of binary complexes has been extensively studied, and it is well understood in cases when rigid-body recognition is a valid approximation (e.g. barnase–barstar and

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

antigen–antibody complexes). In comparison, flexible recognition and the self-assembly of multi-component systems are still enigmatic. Both processes are of great interest to biologists; they are undoubtedly complex, but they deserve that physicists and biophysicists give them the same attention as they have to protein folding over the past 20 years.

6. Acknowledgements

We thank E. Lévy (Cambridge, UK) for Fig. 1 and F. Cazals (Sophia-Antipolis, France) for Fig. 3. We acknowledge a Research Associateship to R.P.B. of the Deutsche Forschungsgemeinschaft, support to J.J. of the 3D-Repertoire and SPINE2-Complexes programs of the European Union, and financial support from DBT, India, to P.C.

7. References

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A. & Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294.

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L. & Russell, R. B. (2004). Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026–2029.

Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology* **332**, 989–998.

Aloy, P., Pichaud, M. & Russell, R. B. (2005). Protein complexes: structure prediction challenges for the 21st century. *Current Opinion in Structural Biology* **15**, 15–22.

Aloy, P. & Russell, R. B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **19**, 161–162.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A. *et al.* (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* **29**, 37–40.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Engineering* **2**, 101–113.

Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface-mapping of phylogenetic Information. *Journal of Molecular Biology* **307**, 447–463.

Aurenhammer, F. (1987). Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing* **16**, 78–96.

Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins* **53**, 708–719.

Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein–protein interfaces. *Journal of Molecular Biology* **336**, 943–955.

Bahadur, R. P. & Janin, J. (2008). Residue conservation in viral capsid assembly. *Proteins* **71**, 407–414.

Bahadur, R. P., Rodier, F. & Janin, J. (2007). A dissection of the protein–protein interfaces in icosahedral virus capsids. *Journal of Molecular Biology* **367**, 574–590.

Bahadur, R. P., Zacharias, M. & Janin, J. (2008). Dissecting protein–RNA sites. *Nucleic Acids Research* **26**, 2705–2716.

Baldwin, J. & Chothia, C. (1979). Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *Journal of Molecular Biology* **129**, 175–220.

Ban, Y. E. A., Edelsbrunner, H. & Rudolph, J. (2004). Interface surfaces for protein–protein complexes. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology* (RECOMB 2004), pp. 205–212. San Diego, CA.

Benesch, J. L. & Robinson, C. V. (2006). Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Current Opinion in Structural Biology* **16**, 245–251.

Berchanski, A., Segal, D. & Eisenstein, M. (2005). Modeling oligomers with C_n or D_n symmetry: application to CAPRI target 10. *Proteins* **60**, 202–206.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242.

Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J. & Poupon, A. (2008). DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

biological protein–protein interactions. *Bioinformatics* **24**, 652–658.

Bhat, T. N., Bentley, G. A., Boulot, G., Greene, M. I., Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F. P., Mariuzza, R. A. & Poljak, R. J. (1994). Bound water molecules and conformational stabilization help mediate an antigen–antibody association. *Proceedings of the National Academy of Sciences USA* **91**, 1089–1093.

Birtalan, S., Zhang, Y., Fellouse, F. A., Shao, L., Schaefer, G. & Sidhu, S. S. (2008). The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *Journal of Molecular Biology* **377**, 1518–1528.

Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C. A. & Klebe, G. (2006). Physicochemical descriptors to discriminate protein–protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins* **65**, 607–622.

Bode, W., Wei, A. Z., Huber, R., Meyer, E., Travis, J. & Neumann, S. (1986). X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO Journal* **5**, 2453–2458.

Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* **280**, 1–9.

Braden, B. C. & Poljak, R. J. (2000). Structure and energetics of anti-lysozome antibodies. In *Protein–protein recognition* (ed. C. Kleanthous), pp. 126–161. Oxford University Press.

Bradley, P., Misura, K. M. & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871.

Bressanelli, S., Stiasny, K., Allison, S. L., Stura, E. A., Duquerroy, S., Lescar, J., Heinz, F. X. & Rey, F. A. (2004). Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO Journal* **23**, 728–738.

Caffrey, D., Somaroo, S., Hughes, J., Mintseris, J. & Huang, E. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science* **13**, 190–202.

Carugo, O. & Argos, P. (1997). Protein–protein crystal-packing contacts. *Protein Science* **6**, 2261–2263.

Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science* **10**, 1470–1473.

Caspar, D. L. (1980). Movement and self-control in protein assemblies. Quasi-equivalence revisited. *Biophysical Journal* **10**, 103–135.

Caspar, D. L. & Klug, A. (1962). Physical principles in the construction of regular viruses. *Cold Spring Harbor Symposia on Quantitative Biology* **27**, 1–24.

Cazals, F., Proust, F., Bahadur, R. P. & Janin, J. (2006). Revisiting the Voronoi description of protein–protein interfaces. *Protein Science* **15**, 2082–2092.

Chakrabarti, P. & Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins* **47**, 334–343.

Chakravarty, S. & Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **7**, 723–732.

Chaudhury, S., Sircar, A., Sivasubramanian, A., Berrondo, M. & Gray, J. J. (2007). Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6–12. *Proteins* **69**, 793–800.

Chen, H. & Skolnick, J. (2008). M-TASSER: an algorithm for protein quaternary structure prediction. *Biophysical Journal* **94**, 918–928.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338–339.

Chothia, C. & Janin, J. (1975). Principles of protein–protein recognition. *Nature* **256**, 705–708.

Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone–receptor interface. *Science* **267**, 383–386.

Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. (2004). ClusPro: a fully automated algorithm for protein–protein docking. *Nucleic Acids Research* **32**, W96–W99.

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713.

Copley, R. R., Ponting, C. P., Schultz, J. & Bork, P. (2002). Sequence analysis of multidomain proteins: past perspectives and future directions. *Advances in Protein Chemistry* **61**, 75–98.

Crick, F. H. & Watson, J. D. (1956). Structure of small viruses. *Nature* **177**, 473–475.

Crosio, M. P., Janin, J. & Jullien, M. (1992). Crystal packing in six crystal forms of pancreatic ribonuclease. *Journal of Molecular Biology* **228**, 243–251.

Crowley, P. B. & Carrondo, M. A. (2004). The architecture of the binding site in redox protein complexes: implications for the fast dissociation. *Proteins* **55**, 603–612.

Darnall, D. W. & Klotz, I. M. (1975). Subunit constitution of proteins: a table. *Archives of Biochemistry and Biophysics* **166**, 651–682.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmström, L., Wollacott, A. M., Wang, C., Andre, I. & Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**, 118–128.

Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. & Bell, J. A. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins* **28**, 494–514.

de Vries, S. J., van Dijk, A. D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T. & Bonvin, A. M. (2007). HADDOCK *versus* HADDOCK: new

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **69**, 726–733.

DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Current Opinion in Structural Biology* **12**, 14–20.

Dokland, T. (2000). Freedom and restraint: themes in virus assembly. *Structure* **8**, R157–R162.

Dokland, T., Bernal, R. A., Burch, A., Pletnev, S., Fane, B. A. & Rossmann, M. G. (1999). The role of scaffolding proteins in the assembly of the small, single-stranded DNA virus phiX174. *Journal of Molecular Biology* **288**, 595–608.

Dominguez, C., Boelens, R. & Bonvin, A. M. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731–1737.

Dupuis, F., Sadoc, J., Jullien, R., Angelov, B. & Mornon, J. P. (2005). Voro3D: 3D Voronoi tesselation applied to protein structures. *Bioinformatics* **21**, 1715–1716.

Dutta, S. & Berman, H. M. (2005). Large macromolecular complexes in the Protein Data Bank: a status report. *Structure* **13**, 381–388.

Edelsbrunner, H. & Mucke, E. P. (1994). Three-dimensional alpha-shapes. *ACM Transactions on Graphics* **13**, 43–72.

Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* **405**, 823–826.

Elcock, A. & McCammon, J. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proceedings of the National Academy of Sciences USA* **98**, 2990–2994.

Ellis, J. J., Broom, M. & Jones, S. (2007). Protein–RNA interactions: structural analysis and functional classes. *Proteins* **66**, 903–911.

Ellis, J. J. & Jones, S. (2008). Evaluating conformational changes in protein structures binding RNA. *Proteins* **70**, 1518–1526.

Finn, R. D., Marshall, M. & Bateman, A. (2005). iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410–412.

Fu, H. (Ed.) (2004). *Protein–Protein Interactions: Methods in Molecular Biology*, vol. 261, 532 pp. Totowa NJ: Humana Press.

Gabb, H. A., Jackson, R. M. & Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology* **272**, 106–120.

Gao, Y., Douguet, D., Tovchigrechko, A. & Vakser, I. A. (2007). DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. *Proteins* **69**, 845–851.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R. *et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636.

Gellatly, B. J. & Finney, J. L. (1982). Calculation of protein volumes: an alternative to the Voronoi procedure. *Journal of Molecular Biology* **161**, 305–322.

Gerstein, M. & Chothia, C. (1996). Packing at the protein–water interface. *Proceedings of the National Academy of Sciences USA* **93**, 10167–10172.

Gerstein, M., Tsai, J. & Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *Journal of Molecular Biology* **249**, 955–966.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B. *et al.* (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736.

Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins* **43**, 89–102.

Goodsell, D. S. & Olson, A. J. (2000). Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* **29**, 105–153.

Gray, J. (2006). High-resolution protein–protein docking. *Current Opinion in Structural Biology* **16**, 150–169.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. & Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology* **331**, 281–299.

Grimes, J. M., Burroughs, J. N., Gouet, P., Diprose, J. M., Malby, R., Ziéntara, S., Mertens, P. P. & Stuart, D. I. (1998). The atomic structure of the bluetongue virus core. *Nature* **395**, 470–478.

Grimm, V., Zhang, Y. & Skolnick, J. (2006). Benchmarking of dimeric threading and structure refinement. *Proteins* **63**, 457–465.

Grünberg, R., Nilges, M. & Leckner, J. (2006). Flexibility and conformational entropy in protein–protein binding. *Structure* **14**, 683–693.

Guharoy, M. & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein–protein interfaces. *Proceedings of the National Academy of Sciences USA* **102**, 15447–15452.

Guharoy, M. & Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics* **15**, 1909–1918.

Hadfield, A. T., Lee, W., Zhao, R., Oliveira, M. A., Minor, I., Rueckert, R. R. & Rossmann, M. G. (1997). The refined structure of human rhinovirus 16 at 2.15 Å resolution: implications for the viral life cycle. *Structure* **5**, 427–441.

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Halperin, I., Wolfson, H. & Nussinov, R. (2004). Protein–protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure* **12**, 1027–1038.

Halperin, I., Wolfson, H. & Nussinov, R. (2006). Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin–Dockerin families. *Proteins* **63**, 832–845.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure* **2**, 641–649.

Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. (1978). Tomato bushy stunt virus at 2·9 Å resolution. *Nature* **276**, 368–373.

Headd, J. J., Ban, Y. E., Brown, P., Edelsbrunner, H., Vaidya, M. & Rudolph, J. (2007). Protein–protein interfaces: properties, preferences, and projections. *Journal of Proteome Research* **6**, 2576–2586.

Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends in Biochemical Sciences* **23**, 358–361.

Hubbard, S. J. & Argos, P. (1994). Cavities and packing at protein interfaces. *Protein Science* **3**, 2194–2206.

Hwang, H., Pierce, B., Mintseris, J., Janin, J. & Weng, Z. (2008). Protein–protein docking benchmark version 3.0. *Proteins* [Epub ahead of print] May 19.

Inbar, Y., Benyamini, H., Nussinov, R. & Wolfson, H. J. (2005). Prediction of multimolecular assemblies by multiple docking. *Journal of Molecular Biology* **349**, 435–447.

Janin, J. (1997). Specific *versus* non-specific contacts in protein crystals. *Nature Structural & Molecular Biology* **4**, 973–974.

Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein–protein and protein–DNA recognition. *Structure Fold Design* **7**, R277–R279.

Janin, J. (2005). Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Science* **14**, 278–283.

Janin, J. (2007). Structural genomics: winning the second half of the game. *Structure* **15**, 1347–1349.

Janin, J. & Chothia, C. (1976). Stability and specificity of protein–protein interactions: the case of the trypsin–trypsin inhibitor complexes. *Journal of Molecular Biology* **100**, 197–211.

Janin, J. & Chothia, C. (1990). The structure of protein–protein recognition sites. Journal of Biological Chemistry **265**, 16027–16030.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I. & Wodak, S. J. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2–9.

Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *Journal of Molecular Biology* **204**, 155–164.

Janin, J. & Rodier, F. (1995). Protein–protein interaction at crystal contacts. *Proteins* **23**, 580–587.

Janin, J., Rodier, F., Chakrabarti, P. & Bahadur, R. P. (2007). Macromolecular recognition in the Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography* **63**, 1–8.

Janin, J. & Wodak, S. J. (Eds.) (2003). *Protein Modules and Protein–Protein Interaction: Advances in Protein Chemistry*, vol. 61, 333 pp. San Diego, London Academic Press.

Janin, J. & Wodak, S. (2007). The third CAPRI assessment meeting. *Structure* **15**, 755–759.

Johnson, J. E. & Speir, J. A. (1997). Quasi-equivalent viruses: a paradigm for protein assemblies. *Journal of Molecular Biology* **269**, 665–675.

Jones, S., Daley, D., Luscombe, N., Berman, H. M. & Thornton, J. M. (2001). Protein–RNA interactions: a structural analysis. *Nucleic Acids Research* **29**, 943–954.

Jones, S. & Thornton, J. M. (1995). Protein–protein interactions: a review of protein dimer structures. *Progress in Biophysics and Molecular Biology* **63**, 31–65.

Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proceedings of the National Academy of Sciences USA* **93**, 13–20.

Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *Journal of Molecular Biology* **272**, 121–132.

Jones, S. & Thornton, J. M. (2000). Analysis and classification of protein–protein interactions from a structural perspective. In *Protein–Protein Recognition, Frontiers in Molecular Biology*, vol. 31 (ed. C. Kleanthous), pp. 33–59. New York: Oxford University Press.

Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *Journal of Molecular Biology* **287**, 877–896.

Juan, D., Pazos, F. & Valencia, A. (2008). Co-evolution and co-adaptation in protein networks. *FEBS Letters* **582**, 1225–1230.

Kelly, C. A., Nishiyama, M., Ohnishi, Y., Beppu, T. & Birktoft, J. J. (1993). Determinants of protein thermostability observed in the 1.9-Å crystal structure of malate dehydrogenase from the thermophilic bacterium *Thermus flavus*. *Biochemistry* **32**, 3913–3922.

Keskin, O., Bahar, I., Badretdinov, A. Y., Ptitsyn, O. B. & Jernigan, R. L. (1998). Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Science* **7**, 2578–2586.

Keskin, O., Ma, B. & Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology* **345**, 1281–1294.

Kleanthous, C., ed. (2000). *Protein–Protein Recognition: Frontiers in Molecular Biology*, 314 pp., New York: Oxford University Press.

Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **372**, 774–797.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A. and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643.

Krol, M., Chaleil, R. A., Tournier, A. L. & Bates, P. A. (2007). Implicit flexibility in protein docking: cross-docking and local refinement. *Proteins* **69**, 750–757.

Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E. & Sigler, P. B. (1996). The 2·0 Å crystal structure of a heterotrimeric G protein. *Nature* **379**, 311–319.

Larsen, T. A., Olson, A. J. & Goodsell, D. S. (1998). Morphology of protein–protein interfaces. *Structure* **6**, 421–427.

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics* **13**, 323–330.

Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *Journal of Molecular Biology* **234**, 946–950.

Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology* **55**, 379–400.

Leiman, P. G., Kanamaru, S., Mesyanzhinov, V. V., Arisaka, F. & Rossmann, M. G. (2003). Structure and morphogenesis of bacteriophage T4. *Cellular and Molecular Life Sciences* **60**, 2356–2370.

Lensink, M. F., Méndez, R. & Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins* **69**, 704–718.

Lévy, E. D. (2007). PiQSi: protein quaternary structure investigation. *Structure* **15**, 1364–1367.

Lévy, E. D., Erba, E. B., Robinson, C. V. & Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature* **453**, 1262–1265.

Lévy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. (2006). 3D complex: a structural classification of protein complexes. *PLoS Computational Biology* **2**, e155.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E. & Vidal, M. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543.

Li, Y., Huang, Y., Swaminathan, C. P., Smith-Gill, S. J. & Mariuzza, R. A. (2005). Magnitude of the hydrophobic effect at central *versus* peripheral sites in protein–protein interfaces. *Structure* **13**, 297–307.

Lichtarge, O., Bourne, H. & Cohen, F. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, **257**, 342–358.

Lichtarge, O. & Sowa, M. (2002). Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology* **12**, 21–27.

Liljas, L. (1999). Virus assembly. *Current Opinion in Structural Biology* **9**, 129–134.

Liljas, L., Unge, T., Jones, T. A., Fridborg, K., Lövgren, S., Skoglund, U. & Strandberg, B. (1982). Structure of satellite tobacco necrosis virus at 3.0 Å resolution. *Journal of Molecular Biology* **159**, 93–108.

Linderström-Lang, K. U. & Schellman, J. A. (1959). Protein structure and enzyme activity. In *The Enzymes*, vol. 1, 2nd edn (eds. Boyer, D., Lardy, H & Myrback, K.), pp. 443–510. New York: Academic Press.

Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *Journal of Molecular Biology* **285**, 2177–2198.

Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. (2003). Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences USA* **100**, 5772–5777.

Marshall, G. R. & Vakser, I. A. (2005). Protein–protein docking methods. In *Proteomics and Protein–Protein Interaction: Biology, Chemistry, Bioinformatics, and Drug Design* (ed. G. Waksman), pp. 115–146. New York: Springer.

May, A. & Zacharias, M. (2007). Protein–protein docking in CAPRI using ATTRACT to account for global and local flexibility. *Proteins* **69**, 774–780.

McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology* **238**, 777–793.

Méndez, R., Leplae, R., De Maria, L. & Wodak, S. J. (2003). Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins* **52**, 51–67.

Méndez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

3–5 shows progress in docking procedures. *Proteins* **60**, 150–169.

Mihalek, I., Res, I. & Lichtarge, O. (2006). Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. *Bioinformatics* **22**, 1656–1657.

Mihalek, I., Res, I. & Lichtarge, O. (2007). On itinerant water molecules and detectability of protein–protein interfaces through comparative analysis of homologues. *Journal of Molecular Biology* **369**, 584–595.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of Molecular Biology* **196**, 641–656.

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. & Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. *Proteins* **69**, 511–520.

Mintseris, J. & Weng, Z. (2005). Structure, function and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences USA* **102**, 10930–10935.

Monod, J., Changeux, J. P. & Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology* **6**, 306–329.

Monod, J., Wyman, J. & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *Journal of Molecular Biology* **12**, 88–118.

Moont, G., Gabb, H. A. & Sternberg, M. J. E. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35**, 364–373.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **247**, 536–540.

Nadassy, K., Tomas-Oliveira, I., Alberts, I., Janin, J. & Wodak, S. J. (2001). Standard atomic volumes in double-stranded DNA and packing of protein–DNA interfaces. *Nucleic Acids Research* **29**, 3362–3376.

Nadassy, K., Wodak, S. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry* **38**, 1999–2017.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296.

Noreen, I. M. & Thornton, J. M. (2003a). Diversity of protein–protein interactions. *EMBO Journal* **22**, 3486–3492.

Noreen, I. M. & Thornton, J. M. (2003b). Structural characterisation and functional significance of transient protein–protein interactions. *Journal of Molecular Biology* **325**, 991–1018.

Ofran, Y. & Rost, B. (2003). Analysing six types of protein–protein interfaces. *Journal of Molecular Biology* **325**, 377–387.

Ofran, Y. & Rost, B. (2007). ISIS: interaction sites identified from sequence. *Bioinformatics* **23**, e13–e16.

Pal, A., Chakrabarti, P., Bahadur, R., Rodier, F. & Janin, J. (2007). Peptide segments in protein–protein interfaces. *Journal of Biosciences* **32**, 101–111. Erratum in *Journal of Biosciences* 2007 **32**, 805.

Pauling, L. & Corey, R. B. (1951). Configuration of polypeptide chains. *Nature* **168**, 550–551.

Pazos, F. & Valencia, A. (2002). *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227.

Perutz, M. F. (1960). Structure of hemoglobin. *Brookhaven Symposia in Biology* **13**, 165–183.

Perutz, M. F. (1970). Stereochemistry of cooperative effects in haemoglobin. *Nature* **228**, 726–739.

Peschard, P., Kozlov, G., Lin, T., Mirza, I. A., Berghuis, A. M., Lipkowitz, S., Park, M. & Gehring, K. (2007). Structural basis for ubiquitin-mediated dimerization and activation of the ubiquitin protein ligase Cbl-b. *Molecular Cell* **27**, 474–485.

Pierce, B., Tong, W. & Weng, Z. (2005). M-ZDOCK: a grid-based approach for C_n symmetric multimer docking. *Bioinformatics* **21**, 1472–1478.

Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**, 47–57.

Ponstingl, H., Kabir, T., Gorse, D. & Thornton, J. M. (2005). Morphological aspects of oligomeric protein structures. *Progress in Biophysics and Molecular Biology* **89**, 9–35.

Ponstingl, H., Kabir, T. & Thornton, J. M. (2003). Automatic inference of protein quaternary structure from crystals. *Journal of Applied Crystallography* **36**, 1116–1122.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of Molecular Biology* **264**, 121–136.

Poole, A. M. & Ranganathan, R. (2006). Knowledge-based potentials in protein design. *Current Opinion in Structural Biology* **16**, 508–513.

Poupon, A. (2004). Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology* **14**, 233–241.

Poupon, A. & Janin, J. (in press). Analysis and prediction of protein quaternary structure. In *Biological Data Mining, Methods in Molecular Biology* (ed. O. Carugo), (In press). Totowa, NJ: Humana Press.

Prevelige Jr., P. E., Thomas, D. & King, J. (1993). Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells. *Biophysical Journal* **64**, 824–835.

Reddy, V. S., Giesing, H. A., Morton, R. T., Kumar, A., Post, C. B., Brooks, C. L. 3rd & Johnson, J. E. (1998). Energetics of quasiequivalence: computational analysis

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

of protein–protein interactions in icosahedral viruses. *Biophysical Journal* **74**, 546–558.

Reichmann, D., Rahat, O., Albeck, S., Meged, R., Dym, O. & Schreiber, G. (2005). The modular architecture of protein–protein binding interfaces. *Proceedings of the National Academy of Sciences USA* **102**, 57–62.

Reichmann, D., Rahat, O., Cohen, M., Neuvirth, H. & Schreiber, G. (2007). The molecular architecture of protein–protein binding sites. *Current Opinion in Structural Biology* **17**, 67–76.

Res, I. & Lichtarge, O. (2005). Character and evolution of protein–protein interfaces. *Physical Biology* **2**, S36–S43.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of Molecular Biology* **82**, 1–14.

Robinson, C. V., Sali, A. & Baumeister, W. (2007). The molecular sociology of the cell. *Nature* **450**, 973–982.

Rodier, F., Bahadur, R. P., Chakrabarti, P. & Janin, J. (2005). Hydration of protein–protein interfaces. *Proteins* **60**, 36–45.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature* **317**, 145–153.

Rossmann, M. G. & Johnson, J. E. (1989). Icosahedral RNA virus structure. *Annual Review of Biochemistry* **58**, 533–573.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M. & Sali, A. (2004). A structural perspective on protein–protein interactions. *Current Opinion in Structural Biology* **14**, 313–324.

Saha, R. P., Bahadur, R. P. & Chakrabarti, P. (2005). Interresidue contacts in proteins and protein–protein interfaces and their use in characterizing the homodimeric interface. *Journal of Proteome Research* **4**, 1600–1609.

Saha, R. P., Bahadur, R. P., Pal, A., Mandal, S. & Chakrabarti, P. (2006). ProFace: a server for the analysis of the physicochemical features of protein–protein interfaces. *BMC Structural Biology* **6**, 11.

Sali, A. (1998). 100 000 protein structures for the biologist. *Nature Structural Biology* **5**, 1029–1032.

Sander, C. & Schneider, R. (1993). The HSSP data base of protein structure–sequence alignments. *Nucleic Acids Research* **21**, 3105–3109.

Sanger, F. & Thompson, E. O. (1953). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* **53**, 353–366.

Sarai, A. & Kono, H. (2005). Protein–DNA recognition patterns and predictions. *Annual Review of Biophysics and Biomolecular Structure* **34**, 379–398.

Scheffzek, K., Ahmadian, M. R., Kabsch, W., Wiesmüller, L., Lautwein, A., Schmitz, F. & Wittinghofer, A. (1997). The Ras–RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* **277**, 333–338.

Schnackenberg, J., Than, M. E., Mann, K., Wiegand, G., Huber, R. & Reuter, W. (1999). Amino acid sequence, crystallization and structure determination of reduced and oxidized cytochrome *c*6 from the green alga *Scenedesmus obliquus*. *Journal of Molecular Biology* **290**, 1019–1030.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. (2005a). Geometry-based flexible and symmetric protein docking. *Proteins* **60**, 224–231.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. (2005b). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research* **33**, W363–W367.

Schreiber, G. & Fersht, A. R. (1993). Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **32**, 5145–5150.

Schreiber, G., Shaul, Y. & Gottschalk, K. E. (2006). Electrostatic design of protein–protein association rates. *Methods in Molecular Biology* **340**, 235–249.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science* **310**, 638–642.

Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology* **10**, 153–159.

Shepherd, C. M., Borelli, I. A., Lander, G., Natarajan, P., Siddavanahalli, V., Bajaj, C., Johnson, J. E., Brooks 3rd, C. L. & Reddy, V. S. (2006). VIPERdb: a relational database for structural virology. *Nucleic Acids Research* **34**, D386–D389.

Smith, G. R. & Sternberg, M. J. (2002). Prediction of protein–protein interactions by docking methods. *Current Opinion in Structural Biology* **12**, 28–35.

Soyer, A., Chomilier, J., Mornon, J. P., Jullien, R. & Sadoc, J. (2000). Voronoi tesselation reveals the condensed matter character of folded proteins. *Physical Review Letters* **85**, 3532–3535.

Stein, A., Russell, R. B. & Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research* **33**, D413–D417.

Steven, A. C., Heymann, J. B., Cheng, N., Trus, B. L. & Conway, J. F. (2005). Virus maturation: dynamics and mechanism of a stabilizing structural transition that leads to infectivity. *Current Opinion in Structural Biology* **15**, 227–236.

Steven, A. C., Trus, B. L., Booy, F. P., Cheng, N., Zlotnick, A., Caston, J. R. & Conway, J. F. (1997). The making and breaking of symmetry in virus capsid assembly: glimpses of capsid biology from cryoelectron microscopy. *FASEB Journal* **11**, 733–742.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press

Stock, D., Gibbons, C., Arechaga, I., Leslie, A. G. & Walker, J. E. (2000). The rotary mechanism of ATP synthase. *Current Opinion in Structural Biology* **10**, 672–679.

Stock, D., Leslie, A. G. & Walker, J. E. (1999). Molecular architecture of the rotary motor in ATP synthase. *Science* **286**, 1700–1705.

Stockley, P. G., Rolfsson, O., Thompson, G. S., Basnak, G., Francese, S., Stonehouse, N. J., Homans, S. W. & Ashcroft, A. E. (2007). A simple, RNA-mediated allosteric switch controls the pathway to formation of a $T=3$ viral capsid. *Journal of Molecular Biology* **369**, 541–552.

Sundberg, E. J. & Mariuzza, R. A. (2002). Molecular recognition in antibody–antigen complexes. *Advances in Protein Chemistry* **61**, 119–160.

Sundberg, E. J., Urrutia, M., Braden, B. C., Isern, J., Tsuchiya, D., Fields, B. A., Malchiodi, E. L., Tormo, J., Schwarz, F. P. & Mariuzza, R. A. (2000). Estimation of the hydrophobic effect in an antigen–antibody protein–protein interface. *Biochemistry* **39**, 15375–15387.

Svedberg, T. (1927). The ultracentrifuge. *Nobel Lectures, Chemistry 1922–1941*, Elsevier Publishing Company, Amsterdam, 1966.

Thorn, K. S. & Bogan, A. A. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**, 284–285.

Thornton, J. M., Edwards, M. S., Taylor, W. R. & Barlow, D. J. (1986). Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO Journal* **5**, 409–413.

Tovchigrechko, A. & Vakser, I. A. (2006). GRAMM-X public web server for protein–protein docking. *Nucleic Acids Research* **34**, W310–W314.

Tovchigrechko, A., Wells, C. A. & Vakser, I. A. (2002). Docking of protein models. *Protein Science* **11**, 1888–1896.

Treger, M. & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA–protein interfaces. *Journal of Molecular Recognition* **14**, 199–214.

Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Science* **6**, 53–64.

Tsai, J. & Gerstein, M. (2002). Calculation of protein volumes: sensitivity analysis and parameter database. *Bioinformatics* **18**, 985–995.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *Journal of Molecular Biology* **290**, 253–266.

Vajda, S., Weng, Z. & DeLisi, C. (1995). Extracting hydrophobicity parameters from solute partition and protein mutation/unfolding experiments. *Protein Engineering* **11**, 1081–1092.

Valdar, W. S. & Thornton, J. M. (2001). Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108–124.

Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology* **12**, 368–373.

van Dijk, A. D., de Vries, S. J., Dominguez, C., Chen, H., Zhou, H. X. & Bonvin, A. M. (2005). Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins* **60**, 232–238.

Wang, C., Schueler-Furman, O., Andre, I., London, N., Fleishman, S. J., Bradley, P., Qian, B. & Baker, D. (2007). RosettaDock in CAPRI rounds 6–12. *Proteins* **69**, 758–763.

Wells, J. A. & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450**, 1001–1009.

Wiehe, K., Peterson, M. W., Pierce, B., Mintseris, J. & Weng, Z. (2007). Protein–protein docking: overview and performance analysis. *Methods in Molecular Biology* **413**, 283–314.

Xie, Z. & Hendrix, R. W. (1995). Assembly *in vitro* of bacteriophage HK97 proheads. *Journal of Molecular Biology* **253**, 74–85.

Xu, Q., Canutescu, A., Obradovic, Z. & Dunbrack Jr., R. L. (2006). ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics* **22**, 2876–2882.

Zhang, Y., Arakaki, A. K. & Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **61**, 91–98.

Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences USA* **101**, 7594–7599.

Zhu, H., Domingues, F. S., Sommer, I. & Lengauer, T. (2006). NOXclass: prediction of protein–protein interaction types. *BMC Bioinformatics* **7**, 27.

Zlotnick, A. (2005). Theoretical aspects of virus capsid assembly. *Journal of Molecular Recognition* **18**, 479–490.

https://doi.org/10.1017/S0033583508004708 Published online by Cambridge University Press