

CHAPTER 1

X-ray Study of Protein–Protein Complexes and Analysis of Interfaces

Joel Janin

*Yeast Structural Genomics, IBBMC UMR 8612 CNRS,
Université Paris-Sud, 91405 Orsay, France
E-mail: joel.janin@u-psud.fr*

Highly efficient procedures to express genes and prepare individual proteins for structural analysis, developed during the first round of the Structural Genomics initiatives world wide, are now being extended to protein complexes and multi-subunit assemblies. These structures are still few in the Protein Data Bank, but one can exploit the abundant information on binary protein–protein complexes and oligomeric proteins to set up appropriate methods of analysis, and derive rules on protein–protein interaction, which will be applicable to larger assemblies when their structures become available.

1.1 Introduction

Following the completion of the first complete genome sequences at the turn of the century, the question was put to structural biologists: can crystallography and NMR provide three-dimensional structures for the products of all these genes? At that time, it was estimated that a set of 10,000 experimental structures, carefully chosen, would cover the space of existing folds; the remainder could be built by homology.¹ Structural Genomics (SG) initiatives were launched in the USA and Japan in the years 2000–2001, with that goal. With the end of 2009, they will have deposited more than 8,000 new structures in the Protein Data Bank (PDB, <http://www.rcsb.org/pdb/statistics/>), and the target of 10,000 will

almost certainly be reached before 2010. But meanwhile, the landscape around has changed greatly. We now realise that the diversity of DNA sequences may be orders of magnitude greater than what was thought when only a few model genomes were known. Many of the new sequences are unrelated to what we have in the databases, and therefore, many protein folds have yet to be discovered. Moreover, it has become clear that most gene products do not exist and function as single entities. Genome-wide studies of protein–protein interaction have demonstrated that cells contain thousands of macromolecular assemblies of all sizes, from simple dimers to objects that comprise tens or hundreds of polypeptide and/or nucleic acid chains.^{2,3} The examples of the ribosome and the nuclear pore show that the whole assembly, not the individual chains, carries the biological function. The structural analysis should, therefore, not be limited to the isolated components.

The number of solved macromolecular assembly structures is still small compared to that of isolated proteins.⁴ In this review, attempts will be described to characterise macromolecular assemblies similar to the systematic studies that SG initiatives performed on single proteins. While these studies are ongoing, we may look at simpler systems for which the PDB offers more examples: protein–protein complexes and homodimeric proteins. Their atomic structures contain a wealth of information on the chemistry and physical chemistry of the non-covalent interactions that allow polypeptide chains recognising each other and self-assembling into a functional macromolecular entity.^{5–9} The methods developed to extract this information, the observations and rules derived from its analysis, will undoubtedly help us to understand the more complex systems when their structure becomes available.

1.2 Preparing Proteins for Structural Studies

The first genome-wide studies of protein–protein interactions were completed at about the same time as the SG initiatives of the first generation. As a result of that coincidence, the second generation of SG initiatives that started in 2005–2006, included several programmes that are concerned with macromolecular assemblies.^{10–11} Thus, the Yeast Structural Genomics, a small-scale pilot-project that we carried out in

Orsay in 2001–2004, is now part of two programmes funded by the European Union, SPINE2-Complexes and 3D-Repertoire (<http://www.spine2.eu>, <http://www.3drepertoire.org>). Both combine high-resolution X-ray/NMR and medium/low resolution cryo-electron microscopy studies (cryo-EM) in order to study multi-component systems; some of their targets, like RNA polymerase or the exosome that degrades mRNA, have a well-established status in biology. Others have just been identified in systematic tandem-affinity purification/mass spectrometry studies. These complexes have no known function, but with yeast, a wealth of genetic and biochemical tools are available to characterise them while the structural analysis is ongoing. Atomic resolution may not be reachable for some of the targets, but useful models can be obtained by docking into the electron density of cryo-EM images, the high-resolution models obtained by X-ray crystallography on some of the components.

All these studies integrate the expertise acquired by labs that were part of the first round of Structural Genomics initiatives to which they owe many of their tools and first of all, efficient methods to produce and analyse recombinant proteins.¹² Figure 1.1 describes the standard procedure that was set up to express and prepare proteins of *Saccharomyces cerevisiae* during the four years of the Yeast Structural Genomics pilot-project.¹³ It comprises three major steps:

1. *Cloning*: We use the PCR reaction to amplify the target sequence in genomic DNA (mostly intron-free in *S. cerevisiae*); the two primer oligonucleotides contain appropriate restriction sites and the 3'-primer codes for a six-histidine tag placed just after the last codon. The PCR products are purified, digested with restriction enzymes and inserted into an expression vector. Their DNA sequence is checked. In *E. coli*, we use vectors derived from the pET plasmid, which place the target gene under control of the highly efficient phage T7 promoter.

2. *Protein Production*: The level of gene expression and the solubility of the target protein are evaluated in small-scale cultures of several *E. coli* strains, each grown at four different temperatures. The conditions that yield the most soluble protein are retained for large-scale production in 1 litre flasks.

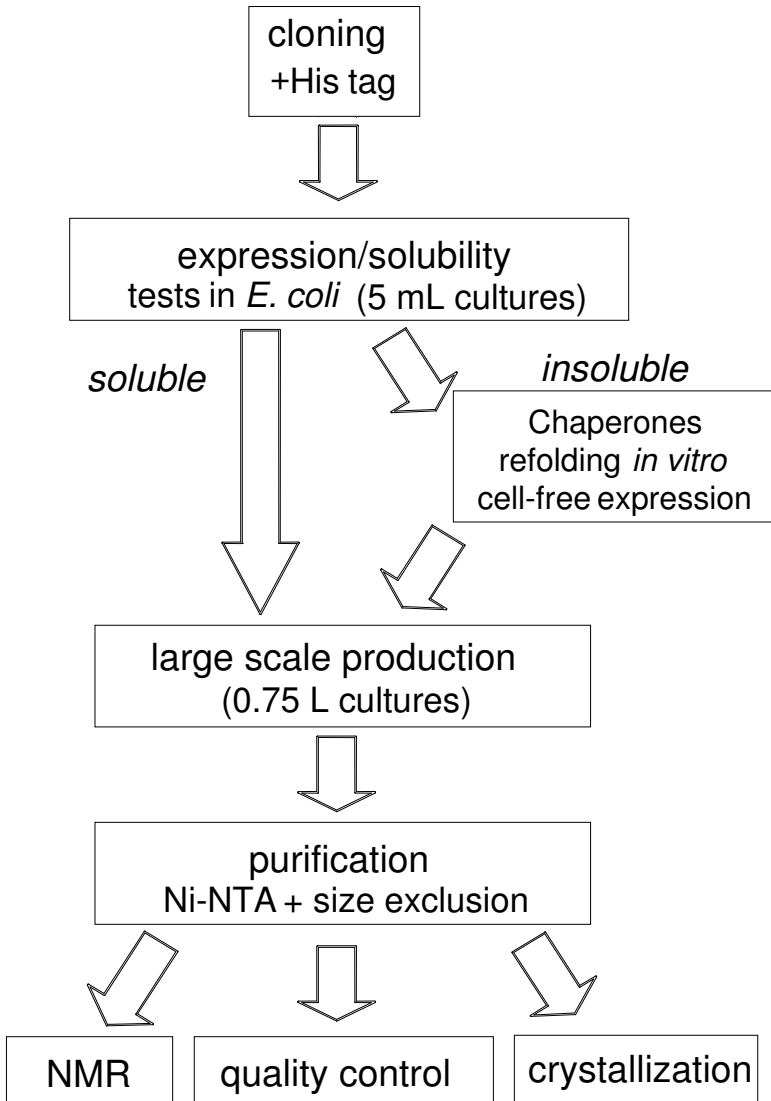


Fig. 1.1. Flowchart of the protein expression/purification procedure. During the Yeast Structural Genomics pilot-project, 250 *S. cerevisiae* genes were cloned and tagged in a standard protein preparation procedure. Expression in *E. coli* succeeded for 80% of the proteins with less than 350 residues. Soluble protein could be purified in two steps from the cell extract, and insoluble protein could be recovered in a number of cases from inclusion bodies (adapted from Ref. 13).

3. Protein Purification and Quality Controls: The His-tagged protein is purified on a Ni-NTA resin, concentrated and run on a size exclusion column. Its degree of purity (usually > 95%) is judged by electrophoresis on a SDS gel and its chemical integrity by mass spectrometry.

The cloning step was carried out on 250 *S. cerevisiae* target genes with a success rate above 90%. After optimization of the growth conditions, most of the cloned genes were highly expressed in *E. coli*; an overnight culture in a shaken flask yielded the target protein in milligram quantities. However, more than one-third of the constructions gave insoluble protein in inclusion bodies. About half of those could be recovered as soluble protein either by co-expressing bacterial chaperones, by solubilizing the inclusion bodies in 6 M guanidinium chloride and screening for refolding in a number of buffers,¹⁴ or by using a cell-free expression system.¹⁵

Carrying out the whole procedure on all the targets was outside the scope of a pilot-project, and therefore, we focused our work on a subset of proteins of interest. Starting with 140 well-expressed yeast genes, we obtained 72 proteins purified to homogeneity in quantities of 0.5 to 10 mg that could be subjected to automated crystallization screens. A majority of the screens gave crystalline hits, not always of sufficient quality for structure determination, but some of these leads could be optimised as discussed below. Fourteen proteins had their X-ray structure determined to resolutions of 1.3 to 2.6 Å within the four-year course of the pilot-project¹⁶ (<http://genomics.eu.org/spip/Overview>), and another ten during the two years after. Therefore, the goal of 20 new structures that we had initially fixed to the pilot-project had been reached by 2006, leaving the place for new projects mostly concerned with protein–protein complexes.

Other SG centres have had success rates similar to ours, often on a much larger scale.¹⁷ The second generation programmes that opened in 2005 in the US and Japan, have built on that experience to set up high-throughput production chains for the structure determination of single gene products by both X-ray crystallography and NMR. Whereas most of the first-generation targets were from prokaryotes or yeast, more difficult targets from higher eukaryotes and including membrane proteins are now being addressed, albeit with a much lower throughput.^{12,18}

1.3 Preparing Protein–Protein Complexes and Multi-component Assemblies

The preparation scheme of Fig. 1.1 has a success rate of 50% that may be considered as satisfactory on a target that is a single gene product. The same scheme can be used to produce multigenic protein assemblies by preparing each component separately. For a binary complex, the expected 25% yield makes it worth trying, but with more than two components the chance is poor that all the subunits can be prepared separately as soluble proteins that will self-assemble when mixed together. Nevertheless, the one-by-one approach has had some remarkable successes. For instance, the *Xenopus* genes coding for the four different histones that constitute the nucleosome core particle could be individually expressed in *E. coli*, and the core particle was reconstituted by mixing them together in appropriate proportions.¹⁹ More frequently, some but not all of the components of a multi-component complex are obtained in soluble form. The complex itself cannot be reconstituted, but some of the soluble components form subcomplexes that can yield important information on the assembly, and they may be suitable for high-resolution structural studies complementing a cryo-EM analysis of the whole complex.

Figure 1.2 describes the strategy that we developed for preparing yeast protein–protein complexes. It offers several alternatives to the one-by-one gene expression approach (Pathway 3). One possible approach is to prepare the assembly directly from yeast extracts, either at its natural abundance (Pathway 1) or after over-expressing all its components (Pathway 2). Over-expression can also be attempted in *E. coli* (Pathway 4). Pathway 1 is the one that was used in the structural studies of bacterial ribosomes, and also of the yeast 20S proteasome.²⁰ The cells can be grown in large quantities, the ribosome and the proteasome are very abundant, and they can be purified by techniques that do not require affinity tags. In all other cases, the complexes must be over-expressed. A simple procedure would be to build an expression vector for each of the genes of interest, and introduce them into the same bacterial or yeast strain. However, it is difficult to maintain more than two plasmids in the same host, and even with a binary complex, the level of expression of

two genes carried by different vectors is likely to be very unequal, compromising the formation of an assembly with a well-defined stoichiometry. The approach that we and others favour is therefore to make operon-like genetic constructions, in which several genes of interest are placed next to each other.

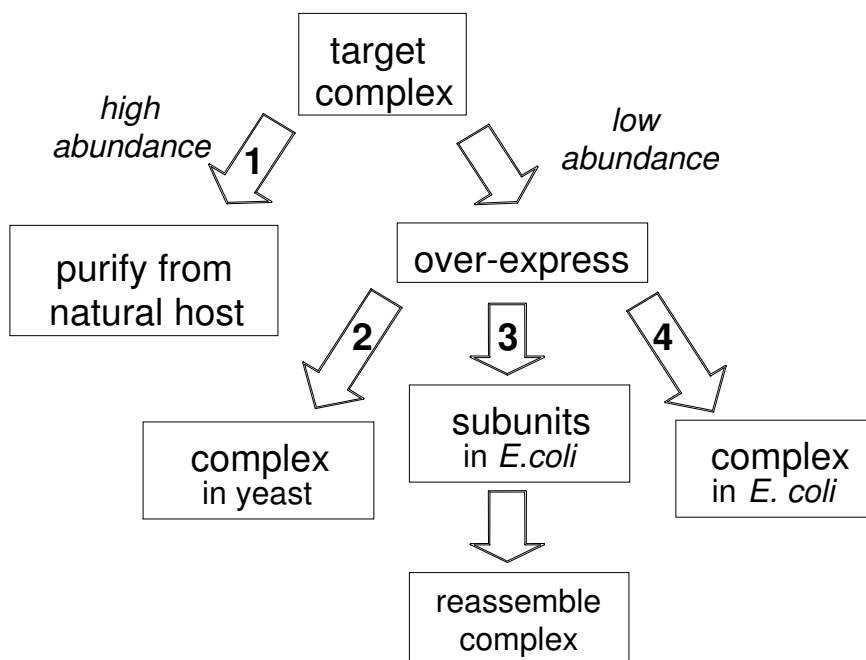


Fig. 1.2. Strategy for the purification of multi-subunit yeast complexes.

They form a single transcription unit under the control of the same promoter, and a ribosome binding site is placed between each stop and start codon.²¹ In practice, five or more medium-size genes can be co-expressed in this way, one of them bearing an affinity purification tag. The construction can be facilitated by placing restriction sites at strategic locations, or dispensed of altogether by using synthetic DNA. The genes

in the operon are transcribed into a single mRNA, they are translated at similar levels and their products are able to associate as they exit the ribosome. Thus, components that would be insoluble (or disordered and degraded) if expressed alone, can be rescued through their interaction with the partner chains. The procedure does not apply to systems such as the complexes of the respiratory chain, because their assembly requires specialized chaperones or cofactors. Still, the co-expression and self-assembly in *E. coli* of eukaryotic protein–protein complexes has had a remarkable success rate, and most of the methodological developments in progress follow Pathway 4.

1.4 Crystallization and X-ray Studies

Crystallization is a well-recognised bottleneck in structural studies. A number of new tools have been developed in recent years, mostly in SG labs. These techniques were designed primarily for single-gene products, but they work equally well for multi-component assemblies and play a key role in the present study. In spite of many attempts to make it rational, the crystallization of proteins, nucleic acids and their complexes still depends on testing hundreds of conditions that combine different precipitants, pHs and additives. One of the very first upshots of the SG initiatives, the one that spread the most quickly, was automatic crystallization. Unlike an attempt we had made²² to use robotics in the early nineties, the devices and procedures that were developed ten years later in the framework of the SG centres immediately found industrial support and are now used routinely by the protein science community. Pipette robots and crystallization kits greatly facilitate the preparation of the precipitant solutions. Equally important, the amount of biological material required to do the tests has dropped by one or two orders of magnitude, thanks to liquid-dispensing robots that prepare arrays of nanodrops in 96-well plates.^{23–24} A standard set of four plates can be prepared in a couple of hours with a minimum of human intervention, and it uses up only a milligram or two of pure protein material. Moreover, the success rate is remarkably high: in our hands, about half of proteins entering crystallization trials give crystals of some sort. As many are not suitable for diffraction experiments because of their size

and shape, or of the low resolution of the X-ray diffraction pattern they give, the conditions under which they are obtained have to be optimised, and that again can be done by automatic procedures.²⁵

The crystallization methods are essentially the same for single proteins and for complexes, although meeting the criterion of homogeneity may be more difficult in the latter case. The protein concentration in the crystallization mix – typically 5–10 mg/ml – allows the formation of low-affinity complexes with K_d values in the micromolar range by simple mixing of the components. But the stoichiometry of the mixture may not be exact and the components in excess can interfere with crystallization, or crystallize separately. It is possible to avoid this problem by purifying the complex on a size exclusion column, or by other chromatographic techniques before crystallization. However, such procedures are only applicable to stable assemblies.

Once diffracting crystals are obtained, the steps that follow usually require labeled material. Labelling is needed for structure solution by both crystallography and NMR. NMR makes extensive use of isotopic labels such as ^{15}N and ^{13}C (see also Chapter 7), and crystallography needs heavy atom labels for phase determination. In a recombinant protein, heavy atom labelling can be very efficiently achieved by incorporating selenomethionine in place of methionine.²⁶ The widespread utilization of the dispersive and anomalous signal of selenium at wavelengths near 0.98 Å has been the key to the development of high-throughput crystallographic methods in the last ten years. Synchrotron radiation centres make this wavelength easily available on experimental setups that allow a complete diffraction dataset to be recorded in a matter of minutes. This is achieved by making full use of the high beam intensity of synchrotron radiation, and by the efficiency of the X-ray detectors developed over the past ten years.

Biological crystallographers have at their disposal an extensive library of software that performs all the steps which follow the recording of X-ray patterns: data reduction, phase determination, model building and refinement.²⁷ Diffraction data taken at several wavelengths (in the Multiple Anomalous Diffraction or MAD method), or even at a single wavelength (in the Single Anomalous Diffraction or SAD method) on

just one crystal, yield good quality phases and electron density maps in which a large part of the polypeptide chain can be traced automatically.²⁸ In the case of a complex, it may be sufficient to label one of the components if prepared separately. On the other hand, selenomethionine incorporation requires the protein to be expressed in bacteria or yeast grown on special media, or possibly *in vivo* in a cell-free system. While the method may be extended to other expression systems in the future, biological material extracted from a natural host cannot be labelled in this way. However, elements other than selenium give a dispersive and anomalous signal that can be used for phasing, for instance the metal ions present in metalloproteins, or just the sulfur atoms of cysteines in favourable cases.²⁹ If none of these methods are applicable, one must return to classical heavy atom labelling techniques; Hg reacting with cysteines for instance, has been used in the past for multiple isomorphous replacement, and is nowadays also suitable for MAD or SAD phasing.

1.5 The Geometric Analysis of Protein–Protein Interfaces

Structure determination by crystallography or NMR ends with the deposition of set atomic coordinates at the Protein Data Bank³⁰ (PDB) that makes it available to the community. The information present in a PDB entry is chemical: the nature of each atom; and geometric: the atomic positions. Its conversion into terms of physical and/or biological relevance is rarely straightforward, and specific tools have been developed for that purpose. Thus, at least three geometric tools are appropriate when defining the interface between two molecules or macromolecules A and B that form a complex AB:

1. *Distance*: atoms or chemical groups i of A and j of B are part of the A:B interface if they satisfy the condition $d_{ij} < d_0$, where 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:

$$d_{ij} < d_0 = r_i + r_j + r_0 \quad (1.1)$$

2. *Buried Surface*:³¹ the A:B interface comprises all the points of the solvent accessible surface of A or B that do not belong to the accessible surface of AB.
3. *The Alpha-complex*,³² a geometric construction related to the Voronoi diagram.

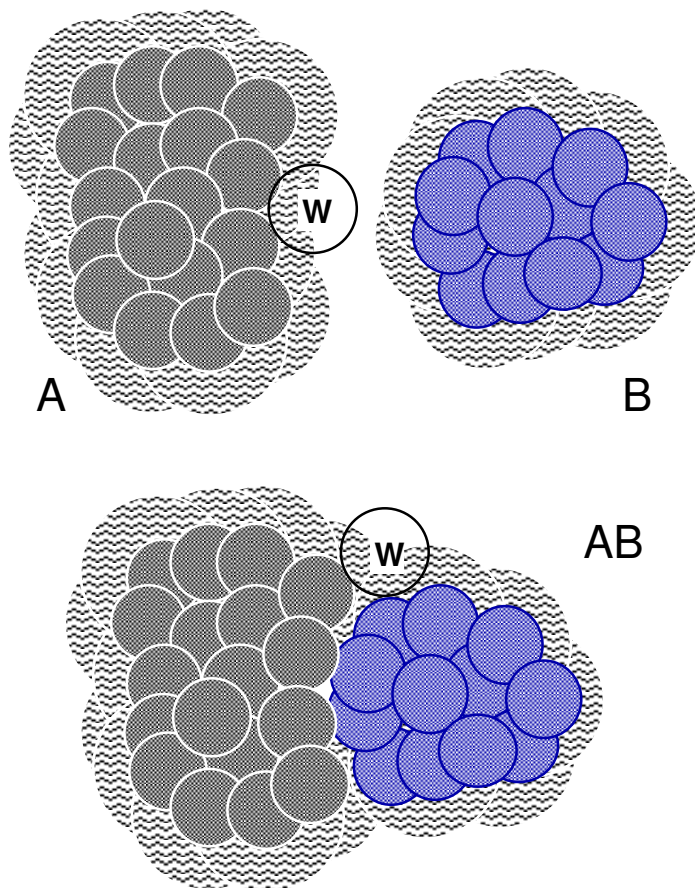


Fig. 1.3. Solvent accessible and buried protein surfaces. The centre of the water probe (W) defines the solvent accessible surface. The A:B interface comprises the points of the solvent accessible surface of A and B (top) that are no longer accessible in the AB complex (bottom).

The first definition is the simplest, but as it depends on an arbitrary cut-off, the second is used more commonly nowadays. In Fig. 1.3, the spheres represent atoms or chemical groups with radii that are augmented of the probe radius ($r_w = 1.4\text{--}1.5 \text{ \AA}$ for a water probe) on the surface; the solvent accessible surface is the border of their union. Its construction may implement the rolling sphere algorithm³³ or an equivalent analytical algorithm. The buried surface area is a convenient measure of the interface size. It can be computed as:

$$\text{BSA} = \text{ASA}_A + \text{ASA}_B - \text{ASA}_{AB} \quad (1.2)$$

where ASA_A , ASA_B and ASA_{AB} are the accessible surface areas of A, B and the AB pair. The atoms and residues that contribute to the BSA are part of the interface, and in practice the BSA is proportional to their number, and also to the number of atom pairs that satisfy Eq. 1 with $r_0 = 2r_w$. The alpha-complex on which the third definition is based is an extension of the Voronoi diagram, a geometric construction first applied to proteins by Richards.³⁴ That diagram associates to each atom its Voronoi cell, the convex polyhedron that contains all points of space closer to that atom than to any other atom. To account for the different sizes of the chemical groups, the Euclidean distance is commonly replaced by the power distance $p(\mathbf{x})$:

$$p(\mathbf{x}) = d^2 - r^2 \quad (1.3)$$

Here, r is the radius of the sphere that represents the atom, and d the distance of point \mathbf{x} to its centre. The Voronoi cell of an atom comprises all points \mathbf{x} that have a power distance to that atom less than to any other atom.^{35–36} Its facets belong to the radical plane, which contains the intersection of the spheres if there is one. The Voronoi or the power diagram offer a natural definition of contacts: two atoms are in contact if their Voronoi cells share a facet. The interface area may then be calculated as the sum of the areas of these ‘bicolour’ facets. In Fig. 1.4a, the blue and red circles that represent atoms of A and B (in two dimensions) have radii that are augmented of r_w as in Fig. 1.3 above. The blue lines are the facets shared by atoms of A, and the green lines are the

bicolour facets shared between A and B. But atoms on the molecular surface always have unbounded facets, like the one between A_1 and A_3 in Fig. 1.4a. They raise a problem to which the alpha-complex gives an elegant solution.³² It 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 spheres do not intersect, or if the facet lies outside the intersection. In Fig. 1.4a, the B_1 and A_3 balls intersect outside their Voronoi cells, due to the presence of A_2 . The corresponding bicolour facet is dashed to indicate that it is not part of the alpha-complex, and the Voronoi interface comprises only the two facets of B_1 with A_1 and A_2 .

An interface defined in this way may still contain a few unbounded facets that Ban *et al.*^{37–38} remove through an iterative retraction procedure, and Cazals *et al.*³⁹ by testing appropriate geometric criteria. Figure 1.4c illustrates how the Voronoi interface defined by the Cazals procedure approximates the shape of the buried molecular surface in the protease–inhibitor complex of Fig. 1.4d. When the two procedures are applied to the set of protein–protein interfaces of Chakrabarti and Janin,⁴⁰ the Voronoi interface area calculated as the sum of the areas of the bicolour facets, correlates linearly with the BSA, but the correlation is much better with the Cazals than the Ban procedure ($R^2 = 0.98$ vs 0.85). A remarkable result of Cazals *et al.*³⁹ is that about 13% of the atoms that share bicolour facets do not contribute to the BSA, mostly because they are not solvent accessible to start with. An example is shown in Fig. 1.4b: on top, the blue and red atoms are shown to share a facet, but the bottom panel indicates that the red atom is not solvent accessible due to the presence of other atoms in molecule B. As a consequence, the Voronoi interface generally comprises significantly more atoms, and especially more main chain atoms, than the buried surface.

1.6 Types and Sizes of Protein–Protein Interfaces

The protein–protein interfaces in the PDB are of several types that represent different categories of interactions.^{41,6–9} One may distinguish between the non-obligate interactions that occur when two preformed

proteins form (non-covalent) complexes, and other interactions that are obligate and permanent.

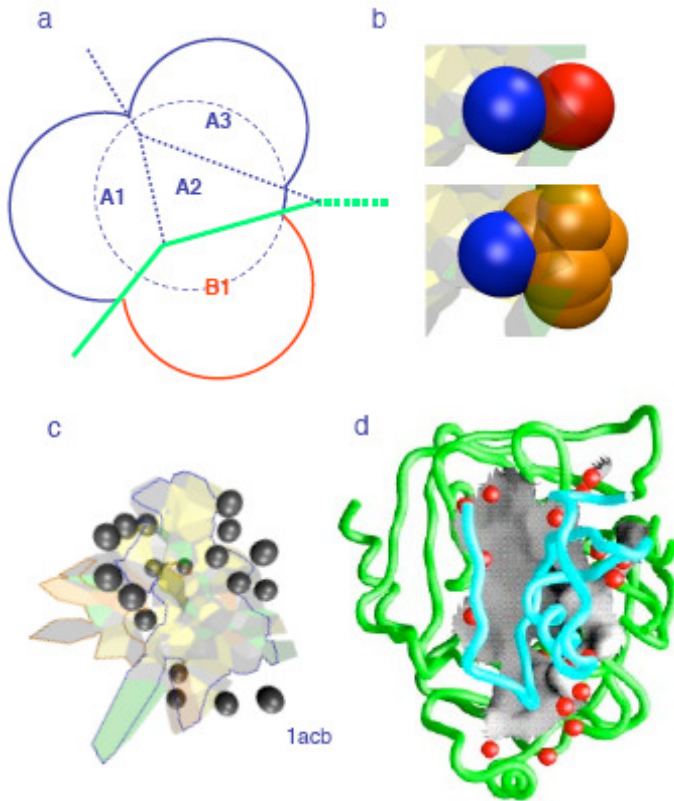


Fig. 1.4. The Voronoi model of protein-protein interfaces. (a) Voronoi interface in two dimensions: The blue and red circles represent atoms of A and B; their radii are augmented of the probe radius; the dashed blue lines are 'monocolour' Voronoi facets shared by atoms of A, the green lines are 'bicolour' facets shared by A and B. The A:B interface comprises the two facets drawn as full lines; they are part of the alpha-complex, but the B1:A3 facet (green dashes) is not. (b) A buried atom can be part of the Voronoi interface: the blue and red balls intersect, although other atoms of molecule B (in gold) make the red ball inaccessible to a solvent probe in the free molecule. (c) The Voronoi interface of a protease-inhibitor complex (1acb); the balls are interface water molecules. (d) The protease surface buried in contact with the inhibitor is viewed through the inhibitor backbone drawn as a blue tube; the green tube is the protease backbone. Panels a-c are adapted from Ref. 39.

The complexes of an antibody with the cognate antigen, of an enzyme with a protein inhibitor, or those that mediate signal transduction in cells, all illustrate non-obligate interactions. In contrast, the interactions between the subunits of an oligomeric protein usually form while their synthesis takes place on the ribosome, or soon after, and break only when the protein is denatured or degraded.

Obligate or not, they play major roles in the structure and function of the protein assemblies that they stabilise. The PDB also contains many examples of a third type of interactions: those that hold protein crystals together. Unlike those that stabilise functional assemblies, crystal packing contacts are unspecific and not subject to any biological selection. They are laboratory artefacts (with some interesting exceptions), yet they are of the same physico-chemical nature as the interactions that stabilise complexes of oligomeric proteins. Any geometric, chemical or physico-chemical feature is of interest if it is able to distinguish between the interfaces created by crystal packing contacts and those that reflect biological interactions, because such a feature may contribute to the specificity of recognition between the protein surfaces involved.

The size of the interface is the most obvious one. Figure 1.5 shows histograms of the BSA in sets of non-obligate protein–protein complexes and homodimeric proteins assembled by Chakrabarti and Janin⁴⁰ and Bahadur *et al.*,⁴² and compares their interfaces with crystal packing interfaces. Mean values and standard deviations are cited in Table 1.1. Sets assembled by Jones *et al.*^{5,43} yield similar values. With the complexes, the distribution peaks near 1,600 Å², and a majority of the interfaces buries less than 2,000 Å². With the homodimers, most of the interfaces are larger, and often much larger. On average, their BSA is twice that of the complexes: 3,900 Å² instead of 1,910 Å², with a large standard deviation which confirms that the sample is very heterogeneous in terms of interface size.

The crystal packing interfaces have a mean BSA of only 570 Å² and therefore, they should be easy to tell apart from the specific interfaces of the complexes and the homodimers. In most cases indeed, a visual inspection of the molecular contacts suffices to identify units of biological relevance, which the PDB calls the ‘biomolecules’.

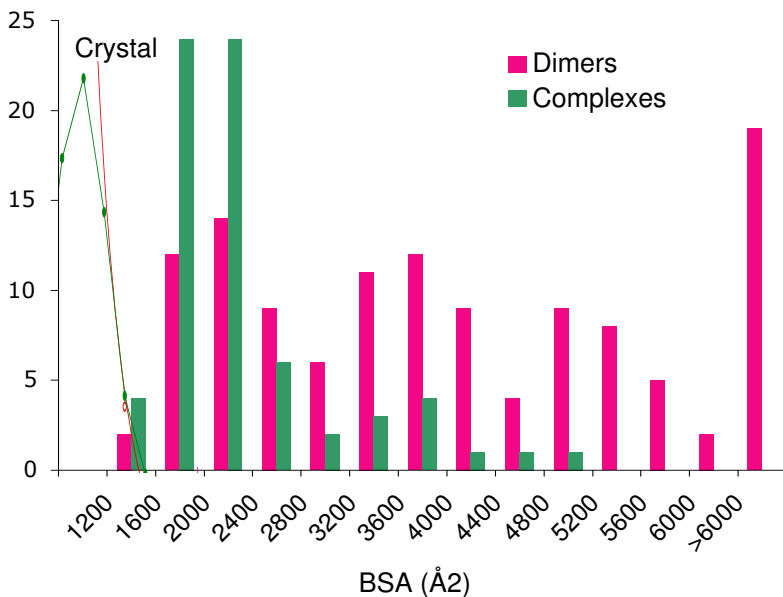


Fig. 1.5. Size distribution of the protein–protein interfaces. Histogram of the BSA in 122 homodimeric proteins, 70 protein–protein complexes, and 1320 crystal packing interfaces (the green line is drawn at a different vertical scale). The references are cited in Table 1.1.

However, what appears in a PDB entry is not the biomolecule, but the crystal asymmetric unit (ASU). The relation between the two is far from obvious: a monomeric protein can yield crystals with two or more chains in the ASU. An oligomeric protein may result in crystals with only one chain, in which case it must have crystal symmetries. A complex may even have subunits in several ASUs. The header of post-1999 PDB entries contains two records, REMARK 300 and REMARK 350, which relate the biomolecule to the content of the ASU. As it takes some effort to convert this information into a set of coordinates, the PDB created Biounit, a database accessible through its RCSB interface, in which the biomolecule is built on the basis of REMARK 300/350 or of supporting information from the authors. The reliability of this procedure can be assessed by comparing it with the composition reported in the

biochemical literature. A wide, albeit incomplete, literature survey carried out by Lévy⁴⁴ indicates that the two disagree in about 15% of the PDB 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) database, well worth checking each time there in case of doubt.

Table 1.1. Properties of protein–protein interfaces.

	Protein–protein complexes ^a	Homodimers ^b	Crystal packing ^c
Number in data set	70	122	188 (1,320)
BSA (\AA^2)	1,910	3,900	1,510 (570)
s.d.	760	2,200	520
Composition (BSA %)			
non-polar	58	65	58
neutral polar	28	23	25
charged	14	12	17
Polar interactions ^d			
number of interface H-bonds	10	19	5
BSA per H-bond (\AA^2)	190	210	280
Water molecules per 1,000 \AA^2	10	11	15
Atomic packing ^e			
Fraction of buried atoms (f_{bu} , %)	34	36	21
Shape complementarity (S_c)	0.69	0.70	0.63
Gap volume index (I_{gap} , \AA)	2.5	2.1	4.4
Packing index (L_D)	42	45	32

a. Data of Chakrabarti and Janin⁴⁰ on a subset of the complexes in Lo Conte *et al.*⁴⁹

b. Data of Bahadur *et al.*⁴²

c. Pairwise interfaces in crystals of monomeric proteins. The values in parentheses are for all the interfaces present in 152 crystal forms of monomeric proteins⁴⁸. All other numbers are for the subset of interfaces with BSA > 800 \AA^2 in Bahadur *et al.*⁴⁷

d. The data on interface water are from Rodier *et al.*⁶¹

e. Data from Bahadur *et al.*⁴⁷

1.7 Chemical and Physical Chemical Properties of the Interfaces

Twelve per cent of the crystal packing interfaces have a BSA above 800 \AA^2 , the size of the smallest interfaces seen in complexes and

homodimers. Thus, they cannot be distinguished from biologically relevant interfaces on the basis of their size only.^{45–47} Moreover, many of the larger crystal packing interfaces are associated with elements of 2-fold symmetry (crystallographic or local). They constitute ‘crystal dimers’ that may be mistaken for real homodimers when interpreting the X-ray structure. Other properties of the interfaces should be considered, their composition for instance. In a protein crystal, each molecule has many neighbours: eight on average in the set of Janin and Rodier;⁴⁸ even though each interface may be small, together they bury a large fraction of the protein surface. The chemical and amino acid composition of this surface must be like that of the solvent accessible surface, whereas the interfaces of complexes and oligomeric proteins are expected to be different. The chemical composition of the protein surface or an interface may be estimated as the fractional contribution of each atom type to the ASA or the BSA. On average, the non-polar (carbon containing) groups, which form 57% of the ASA, contribute marginally more to the BSA in complexes and crystal packing interfaces, but only the interfaces of homodimers are significantly more hydrophobic. The charged groups of Asp/Gly/Lys/Arg side chains are abundant (17–19%) on the protein surface and at crystal packing interfaces relative to the interfaces of complexes or homodimers (12–14%). On average, the crystal packing interfaces contain fewer H-bonds in proportion to their size than in complexes and homodimers: on average, one per 280 Å² of BSA instead of one per 190–210 Å². In contrast, they contain more residual hydration water (Table 1.1). However, the surface composition, the number of H-bonds and the hydration of the interfaces vary widely from a protein or a complex to another.⁴⁹ Thus, the differences between the mean values are always less than the standard deviations.

Still, the procedures that aim to distinguish between biological vs crystal packing interfaces often rely on criteria derived from the hydrophilic/hydrophobic character of the interfaces, in addition to their size. Examples are PQS (Probable Quaternary Structure)⁵⁰ and the related databases of the European Bioinformatics Institute (Hinxton, UK). PQS applies crystal symmetries to the molecules in the ASU, generates neighbours, scores each pairwise interface on the basis of the BSA and a solvation energy term in PQS, and builds molecular

assemblies iteratively by retaining the interfaces that achieve high scores. PITA (Protein InTerfaces and Assemblies)⁵¹ is the same, except that a statistical potential replaces the solvation energy. PISA (Protein Interfaces, Surfaces and Assemblies)⁵¹ uses a graph exploration algorithm to survey all the assemblies that can be formed in the crystal, and an empirical energy is calculated for each. All these procedures ignore the REMARK 300/350 information, and PQS disagrees with it in 18% of the cases.⁵² Moreover, PISA fails to recognise as stable assemblies some classical complexes, the D1.3 antibody–lysozyme complex (1vfb) for instance. Thus, the combined criteria of the interface size and its chemistry are not always sufficient to determine the nature of the biomolecule in a crystal.

Current approaches are based on statistical pairwise potentials and machine-learning procedures (reviewed in several chapters of this volume) that allow many other criteria to be taken into account. They should perform better than PQS or PISA, and in fact they can achieve a success rate close to 95% on test sets of limited size.⁵³ However, they have not yet been applied to the whole PDB, or if they have the results have yet to be made accessible like PQS's or PISA's. In the same way, methods based on phylogeny and sequence conservation (see Chapter 5) have proved their efficiency in a number of cases, but the field of application is still limited.

1.8 Atomic Packing and Interface Topology

Upon visual inspection, the crystal packing interfaces often seem poorly packed and split into small groups of atoms.⁴⁷ In contrast, protein–protein complexes have interfaces that are close-packed like the protein interior⁴⁹ and they form a single contiguous patch,⁴⁰ at least when their BSA is less than 2,000 Å². The quality of the atomic packing and the connectivity of an interface express the shape complementarity of the surfaces in contact. They are important characteristics that govern the energetics of the van der Waals and hydrophobic interactions, but they are not easy to quantify. The volumes of the Voronoi cells can be accurately measured to show that the packing density is the same within 1–2% inside globular proteins⁵⁴ and at the interfaces of protein–protein

complexes⁴⁹ or oligomeric proteins.⁵⁵ But Voronoi volumes can only be estimated for buried atoms (atoms with zero ASA), and those are a minority at macromolecular interfaces: 34–36% on average in complexes and homodimers, 21% at crystal packing interfaces (Table 1.1).

Other descriptors of the geometric complementarity are the S_c index of Lawrence and Colman,⁵⁶ the gap volume index of Laskowski,⁵⁷ and the L_D index of Bahadur *et al.*⁴⁷ The mean values reported in Table 1.1 confirm that the complementarity is less good at crystal packing interfaces than in complexes or homodimers. In Figure 1.6, these values are normalised to 1 for the homodimer interfaces, and their standard deviations are marked. The three descriptors behave similarly, but the contrast is poor with S_c , and the gap volume index has a large standard deviation due in part to the strong edge effects that affect it when the interface is small or split.

A loosely packed interface with a large gap between the two surfaces in contact must bury few atoms in proportion to its size, and therefore the fraction of buried atoms (f_{bu}) is related to the packing. In practice, f_{bu} is at least a good criterion to distinguish specific from non-specific interfaces as S_c or the gap volume index.^{8–9} Interfaces that are split into several regions also bury fewer atoms. Chakrabarti and Janin⁴⁰ applied a geometric clustering algorithm to the interface atoms in order to identify connected regions and define the interface topology. These regions, called recognition patches, generally occur in pairs, one on each protein. A majority of the protein–protein complexes has only one pair, the average number being 1.4. Homodimers, which have larger interfaces, contain more: 1.7 pairs on average, with still a majority of single-pair interfaces.⁴² The algorithm gives unreliable results with crystal packing interfaces. On the other hand, the model based on the alpha-complex provides a straightforward definition of an interface topology: the set of Voronoi facets that constitute the interface can be split into connected components, subsets of facets that have an edge in common, and these subsets correlate well with the recognition patches defined by the clustering algorithm.³⁹

Interfaces may be split in many other ways: along the amino acid sequence into interface chain segments^{5,58} or secondary structure elements.⁵⁹ Chakrabarti and Janin⁴⁰ distinguish within each interface

between a core made of the residues that contain atoms buried in the contact, and a rim in which all interface atoms are solvent accessible. In protein–protein complexes and homodimers, the interface rims have essentially the same amino acid composition as the solvent accessible protein surface, but the cores are significantly different; similarly, the interface core residues tend to be conserved in evolution, but not the rim residues.⁶⁰

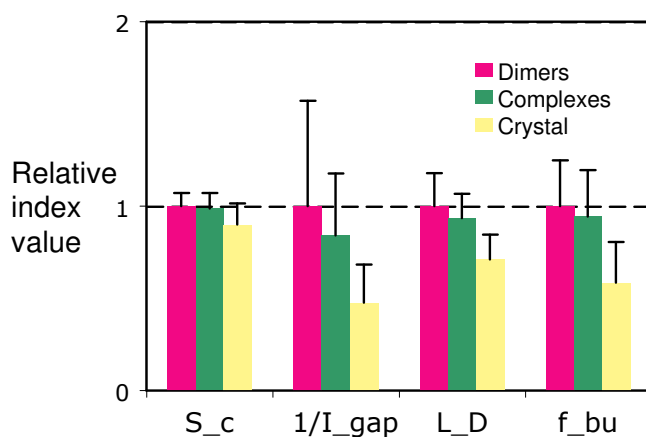


Fig. 1.6. Atomic packing and surface complementarity. Mean values and standard deviations (black bars) of the four parameters reported in Table 1.1 for 122 homodimeric proteins, 70 protein–protein complexes, and 188 crystal packing interfaces with a BSA > 800 Å². All have been scaled to 1 for the homodimer interfaces. Comparatively low values of the S_c shape complementarity index, of the reciprocal of the I_{gap} gap volume index, of the L_D packing index, and of the f_{bu} fraction of buried atoms, all confirm that the atomic packing is less compact and the surface complementarity less good in crystal packing interfaces.

1.9 Conclusions and Outlook

The interfaces of binary assemblies such as the protein–protein complexes and the homodimeric proteins have properties that multi-component assemblies are likely to share, but only to some extent. In

larger oligomeric proteins or in virus capsids,⁶² some of the interfaces are comparable in size, chemical composition and atomic packing density to those of the homodimers. Others are much larger and may span more than two subunits, or they are much smaller and resemble crystal packing interfaces. Presumably, the larger interfaces play a greater role in the stability of the assembly, and are more subject to a selection pressure than small interfaces. The analysis of how different types of interfaces cooperate to stabilise large macromolecular system and contribute to their self-assembly will certainly be one of the most interesting aspects of the ambitious ongoing structural studies.

Acknowledgements

This work greatly benefited from the expertise of Dr A. Poupon, S. Quevillon-Chérueil and other members of the Yeast Structural Genomics team in Orsay, and of B. Séraphin (CNRS, Gif-sur-Yvette). Discussions with F. Cazals (INRIA, Sophia-Antipolis), S. Wodak (University of Toronto), P. Chakrabarti (Bose Institute, Calcutta) and R.P. Bahadur (Jacobs University Bremen), and support of the 3D-Repertoire and SPINE2-Complexes programmes of the European Union are gratefully acknowledged.

References

1. Sali A. (1998). *Nature Struct Biol* 5: 1,029–1,032.
2. Janin J., Séraphin B. (2003). *Curr Opin Struct Biol* 13: 383–388.
3. Devos D., Russell R.B. (2007). *Curr Opin Struct Biol* 17: 370–377.
4. Dutta S., Berman H.M. (2005). *Structure* 13: 381–388.
5. Jones S., Thornton J.M. (1996). *Proc Natl Acad Sci USA* 93: 13–20.
6. Nooren I.M. (2003). *EMBO J* 22: 3,486–3,492.
7. Wodak S.J., Janin J. (2002). *Adv Protein Chem* 61: 9–73.
8. Janin J., Rodier F., Chakrabarti P., Bahadur R.P. (2007). *Acta Crystallogr D Biol Crystallogr* 63: 1–8.
9. Janin J., Bahadur R.P., Chakrabarti P. (2008). *Quart Rev Biophysics* 41: 1–48.
10. Janin J. (2007). *Structure* 15: 1,347–1,349.
11. Vakser I.A. (2008). *Structure* 16: 1–3.
12. Structural Genomics Consortium *et al.* (2008). *Nature Methods* 5: 135–46.

13. Quevillon-Cheruel S., Collinet B., Trésaugues L., Minard P., Henckes G., Aufrère R., Blondeau K., Zhou C.Z., Liger D., Bettache N., Poupon A., Aboulfath I., Leulliot N., Janin J., Van Tilbeurgh H. (2007). Cloning, production, and purification of proteins for a medium-scale structural genomics project. *Macromolecular Crystallography Protocols*, Vol. 1, S. Doublié (ed.) *Methods Mol Biol* 363: 21–37.
14. Trésaugues L., Collinet B., Minard P., Henckes G., Aufrère R., Blondeau K., Liger D., Zhou C.Z., Janin J., Van Tilbeurgh H., Quevillon-Cheruel S. (2004). *J Struct Funct Genomics* 5: 195–204.
15. Kigawa T., Muto Y., Yokoyama S. (1995). *J Biomol NMR* 6: 129–34.
16. Quevillon-Cheruel S., Liger D., Leulliot N., Graille M., Poupon A., de La Sierra-Gallay I.L., Zhou C.Z., Collinet B., Janin J., Van Tilbeurgh H. (2004). *Biochimie* 86: 617–623.
17. Burley S. K., Joachimiak A., Montelione G.T., Wilson I.A. (2008). *Structure* 16: 5–11.
18. Sauder M.J., Rutter M.E., Bain K., Rooney I., Gheyi T., Atwell S., Thompson D.A., Emtage S., Burley S.K. (2008). *Methods Mol Biol* 426: 561–575.
19. Luger K., Rechsteiner T.J., Flaus A.J., Wayne M.M., Richmond T.J. (1997). *J Mol Biol* 272: 301–311.
20. Groll M., Ditzel L., Löwe J., Stock D., Bochtler M., Bartunik H.D., Huber R. (1997). *Nature* 386: 463–471.
21. Tan S. (2001). *Protein Expression and Purification* 21: 224–234.
22. Sadaoui N., Janin J., Lewit-Bentley A. (1994). *J Appl Cryst* 27: 622–626.
23. Bodenstaff E.R., Hoedemaeker F.J., Kuil M.E., de Vrind H.P., Abrahams J.P. (2002). *Acta Crystallogr D Biol Crystallogr* 58: 1,901–1,906.
24. Sulzenbacher G., Gruez A., Roig-Zamboni V., Spinelli S., Valencia C., Pagot F., Vincentelli R., Bignon C., Salomoni A., Grisel S., Maurin D., Huyghe C., Johansson K., Grassick A., Roussel A., Bourne Y., Perrier S., Miallau L., Cantau P., Blanc E., Genevois M., Grossi A., Zenatti A., Campanacci V., Cambillau C. (2002). *Acta Crystallogr D Biol Crystallogr* 58: 2,109–2,115.
25. Leulliot N., Trésaugues L., Bremang M., Sorel I., Ulryck N., Graille M., Aboulfath I., Poupon A., Liger D., Quevillon-Cheruel S., Janin J., Van Tilbeurgh H. (2005). *Acta Crystallogr D Biol Crystallogr* 61: 664–670.
26. Doublié S. (1997). *Methods Enzymol* 276: 523–530.
27. Rossmann M.G., Arnold E. (eds) (2006). Crystallography of Biological Macromolecules. *International Tables for Crystallography* Vol. F.
28. Smith J.L., Hendrickson W.A., Terwilliger T.C., Berendzen J. (2006). Crystallography of Biological Macromolecules. Rossmann M.G. and Arnold E. (eds) *International Tables for Crystallography* Vol. F: 299–309.
29. Dauter Z., Dauter M., de La Fortelle E., Bricogne G., Sheldrick G.M. (1999). *J Mol Biol* 289: 83–92.
30. Berman H.M., Battistuz T., Bhat T.N., Bluhm W.F., Bourne P.E., Burkhardt K., Feng Z., Gilliland G.L., Iype L., Jain S., Fagan P., Marvin J., Padilla D.,

- Ravichandran V., Schneider B., Thanki N., Weissig H., Westbrook J.D., Zardecki C. (2002). *Acta Crystallogr D Biol Crystallogr* 58: 899–907.
31. Chothia C., Janin J. (1975). *Nature* 256: 705–708.
 32. Edelsbrunner H., Mucke E.P. (1994). *ACM Trans Graphics* 13: 43–72.
 33. Lee B.K., Richards F.M., (1971). *J Mol Biol* 55: 379–400.
 34. Richards F.M. (1974). *J Mol Biol* 82: 1–14.
 35. Gellatly B.J., Finney J.L. (1982). *J Mol Biol* 161: 305–322.
 36. Aurenhammer F. (1987). *SIAM J Computing* 16: 78–96.
 37. Ban Y.E.A., Edelsbrunner H., Rudolph J. (2004). *RECOMB* 2: 205–212.
 38. Headd J.J., Ban Y.E., Brown P., Edelsbrunner H., Vaidya M., Rudolph J. (2007). *J Proteome Res* 6: 2,576–2,586.
 39. Cazals F., Proust F., Bahadur R.P., Janin J. (2006). *Protein Sci* 15: 2,082–2,092.
 40. Chakrabarti P., Janin J. (2002). *Proteins* 47: 334–343.
 41. Larsen T.A., Olson A.J., Goodsell D.S. (1998). *Structure* 6: 421–427.
 42. Bahadur R.P., Chakrabarti P., Rodier F., Janin J. (2003). *Proteins* 53: 708–719.
 43. Jones S., Thornton J.M. (1995). *Prog Biophys Mol Biol* 63: 31–65.
 44. Lévy E.D. (2007). *Structure* 1: 1,364–1,375.
 45. Pongstingl H., Henrick K., Thornton J.M. (2000). *Proteins* 41: 47–57.
 46. Pongstingl H., Kabir T., Thornton J.M. (2003). *J Appl Cryst* 36: 1,116–1,122.
 47. Bahadur R.P., Chakrabarti P., Rodier F., Janin J. (2004). *J Mol Biol* 336: 943–955.
 48. Janin J., Rodier F. (1995). *Proteins* 23: 580–587.
 49. Lo Conte L., Chothia C., Janin J. (1999). *J Mol Biol* 285: 2,177–2,198.
 50. Henrick K., Thornton J.M. (1998). *Trends Biochem Sci* 23: 358–361.
 51. Krissinel E., Henrick K. (2007). *J Mol Biol* 372: 774–797.
 52. Xu Q., Canutescu A., Obradovic Z., Dunbrack R.L. Jr (2006). *Bioinformatics* 22: 2,876–2,882.
 53. Bernauer J., Bahadur R.P., Rodier F., Janin J., Poupon A. (2008). *Bioinformatics* 24: 652–658.
 54. Tsai J., Taylor R., Chothia C., Gerstein M. (1999). *J Mol Biol* 290: 253–266.
 55. Pongstingl H., Kabir T., Gorse D., Thornton J.M. (2005). *Prog Biophys Mol Biol* 89: 9–35.
 56. Lawrence M.C., Colman P.M. (1993). *J Mol Biol* 234: 946–950.
 57. Laskowski R.A. (1995). *J Mol Graph* 13: 323–330.
 58. Pal A., Chakrabarti P., Bahadur R., Rodier F., Janin J. (2007). *J Biosci* 32: 101–111.
 59. Guharoy M., Chakrabarti P. (2007). *Bioinformatics* 15: 1,909–1,918.
 60. Guharoy M., Chakrabarti P. (2005). *Proc Nat Acad Sci USA* 102: 15,447–15,452.
 61. Rodier F., Bahadur R.P., Chakrabarti P., Janin J. (2005). *Proteins* 60: 36–45.
 62. Bahadur R.P., Rodier F., Janin J. (2007). *J Mol Biol* 367: 574–590.