Bioinformatics, 32(4), 2016, 481–489
doi: 10.1093/bioinformatics/btv622
Advance Access Publication Date: 27 October 2015
Review

Structural bioinformatics

Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts

Guido Capitani[1,2,*], Jose M. Duarte[1,2], Kumaran Baskaran[1], Spencer Bliven[1,3,4] and Joseph C. Somody[1,5]

[1]Laboratory of Biomolecular Research, Paul Scherrer Institute, OFLC/110, 5232 Villigen PSI, [2]Department of Biology, ETH Zurich, 8093 Zurich, Switzerland, [3]Bioinformatics and Systems Biology Program, UC San Diego, La Jolla, CA 92093, [4]National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA and [5]Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 17, 2015; revised on October 13, 2015; accepted on October 16, 2015

Abstract

Modern structural biology still draws the vast majority of information from crystallography, a technique where the objects being investigated are embedded in a crystal lattice. Given the complexity and variety of those objects, it becomes fundamental to computationally assess which of the interfaces in the lattice are biologically relevant and which are simply crystal contacts. Since the mid-1990s, several approaches have been applied to obtain high-accuracy classification of crystal contacts and biological protein–protein interfaces. This review provides an overview of the concepts and main approaches to protein interface classification: thermodynamic estimation of interface stability, evolutionary approaches based on conservation of interface residues, and co-occurrence of the interface across different crystal forms. Among the three categories, evolutionary approaches offer the strongest promise for improvement, thanks to the incessant growth in sequence knowledge. Importantly, protein interface classification algorithms can also be used on multimeric structures obtained using other high-resolution techniques or for protein assembly design or validation purposes. A key issue linked to protein interface classification is the identification of the biological assembly of a crystal structure and the analysis of its symmetry. Here, we highlight the most important concepts and problems to be overcome in assembly prediction. Over the next few years, tools and concepts of interface classification will probably become more frequently used and integrated in several areas of structural biology and structural bioinformatics. Among the main challenges for the future are better addressing of weak interfaces and the application of interface classification concepts to prediction problems like protein–protein docking.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: guido.capitani@psi.ch

1 Introduction

Contemporary structural biology is a mature yet dynamic field, where well-established techniques like protein crystallography and nuclear magnetic resonance coexist and cross-fertilize with emerging or re-emerging ones, such as electron diffraction or single-particle electron cryomicroscopy. Thanks to advances in detectors and software, this latter technique has broken resolution barriers that were once thought unassailable (Bartesaghi *et al.*, 2015; Campbell *et al.*, 2015; Nogales and Scheres, 2015). The bulk of biomacromolecular structural information; however, still comes from protein crystallography, which accounts for 89% of the entries in the Protein Data Bank (PDB) (Berman, 2000) as of June 2015. A key feature of the crystal structure of biomacromolecules is that the molecules are embedded in a crystal lattice, containing several non-biological interfaces called crystal packing contacts (or, briefly, crystal contacts), which are often indistinguishable by crystallographic means from any biological interface the protein may possess (Fig. 1). With the increasing complexity of biomacromolecular structures, the interface problem—correctly classifying all contacts in a crystal lattice as biologically relevant or crystal contact—has become more frequent and important. Starting in the mid-1990s, several computational approaches, based on a variety of concepts, have been developed to tackle this 'interface classification problem'. The number and diversity of the scientific contributions in this area are now large enough to constitute a recognized topic in structural biology and structural bioinformatics. This review aims to provide an overview of the protein interface classification problem, as well as a historical perspective of the research in this area, of its applications and of the main perspectives and challenges for the future.

2 Determining and annotating the oligomeric state of proteins

Experimentally determining the oligomeric state of a protein in solution can be a difficult task. The determination can be carried out by using a range of biophysical techniques with various degrees of applicability, accuracy and resolution. A non-exhaustive list of such techniques is given in Supplementary Table S1. Particularly challenging is the case of detergent-solubilized transmembrane proteins (TMPs), where the presence of the detergent interferes with the measurements and special measures have to be taken to achieve an accurate molecular mass determination. In many occasions, the outcome of a particular technique might not be conclusive enough and validation by different techniques is desirable. An inherent difficulty with many of these techniques is their limited accuracy in molecular mass determination. At the same time, many of them will not provide the precise location of the biological interfaces, but rather just a global stoichiometry or an approximate idea of the arrangement of molecules. Not infrequently, in the end, linking these experimental data to the interfaces observed in a crystal structure can still lead to ambiguities.

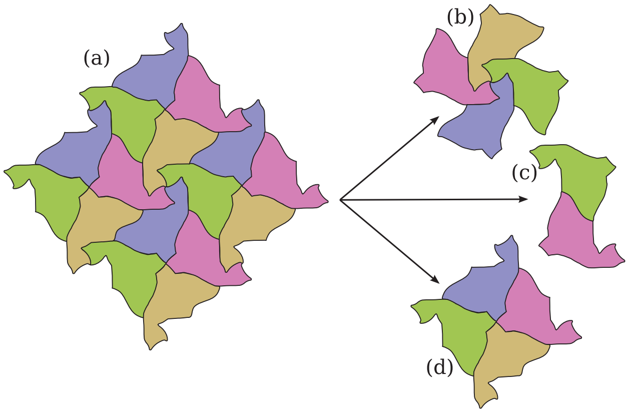

Fig. 1. 2D schematic illustration of the interface classification problem. Given **(a)** as a crystal lattice, any of **(b–d)** could be inferred with equal validity as the biological unit. Without further information, it is not clear which arrangement represents the true biological unit. This figure was inspired by a similar one by Levy and Teichmann (2013), which, in turn, was inspired by 'Fish (N° 20)' by M.C. Escher

Crystallographers are requested upon depositing a new structure to define the biological assembly, with what they believe to be the most likely solution state of the molecule or complex. The PDB stores these 'biological units' in the *pdbx_struct_assembly* category of the mmCIF dictionary (Westbrook and Bourne, 2000) (corresponding to REMARK 350 of the PDB file format) and shows them prominently in their websites (Rose *et al.*, 2011). However, due to missing or ambiguous experimental data and to mistakes during data deposition, these manual annotations are not always reliable. Estimates of the error rate in biological unit annotation in the PDB indicate that the problem is not negligible: Levy found errors in 14.4% of biological units (Levy, 2007), while a study of all protein–protein contacts in the PDB (Baskaran *et al.*, 2014) reported a lower bound of 6.6% of misannotated protein interfaces. In recent years, two independent efforts addressed these difficulties by providing community-based annotation platforms for quaternary structure: PiQSi (Levy, 2007) and PDBWiki (Stehr *et al.*, 2010).

An additional difficulty related to the oligomeric state of a protein is the fact that the complexes exist in an equilibrium, characterized by a certain dissociation constant K_{d}, with values typically in the nanomolar range for strong interactions and in the micromolar or even millimolar range for very weak ones (Ali and Imperiali, 2005). It is important to note that the weakest complexes have free energies of dissociation comparable to those of crystal lattice contacts, so crystallography may not be able to capture them in some cases (Krissinel, 2010). The stability and oligomeric state of complexes is at the same time influenced by the environmental conditions, like pH, ionic strength or temperature. A well-studied example is GAD1 from *Arabidopsis thaliana*, which is a hexamer at low pH (below 6.8) and a dimer at neutral pH (Astegno *et al.*, 2015; Gut *et al.*, 2009). When discussing biological assemblies, we generally consider only complexes that are strongly bound under physiological or close-to-physiological conditions, that is, with a K_d at least in the low micromolar range or smaller. Weak and transient interactions, however, may still be present in protein structures and can lead to ambiguous assemblies.

Defining the biological assembly can be further complicated by the presence of partial structures or partial complexes: in the former case, entire domains or segments of the full-length sequence are deleted for protein preparation or crystallization purposes, while, in the latter, the components of the real assembly are missing. In those cases, strong biological interfaces are likely to be conserved, but the overall biological assembly will be incomplete and may be influenced by crystal contacts. The determination of the oligomeric state should be carried out for the construct used for crystallization. Otherwise, there is no guarantee that a certain oligomeric state of a full-length protein will be the same as that of a single domain extracted from the protein.

It is in response to the above experimental and annotation complexities and difficulties that the field of interface classification in protein crystals started to emerge.

3 Background and history of the field

The first decades of protein crystallography, from the late 1950s to the 1980s, were characterized by structural studies of proteins that could be purified in large quantities from natural sources, that were biochemically well-characterized and that were nearly always soluble and globular, with a well-known quaternary structure. When a new protein structure was finally obtained, after painstaking effort, the preexisting biochemical knowledge made it easy for researchers to find out which protein–protein contacts in the crystal lattice, if any, were biologically relevant and, thus, contributors to the quaternary assembly. The technological advances of the following years, with the mass adoption of recombinant protein production, bright synchrotron radiation sources and much improved structural solution and refinement software, brought about deep changes in the field. Very challenging systems, including multiprotein complexes and TMPs, could now be recombinantly produced, crystallized and structurally solved, sometimes even before a full biochemical analysis of their quaternary structure was available. In fact, the average number of contacts in crystal structures has nearly doubled in the last 30 years (Fig. 2) and is now about 10. As a consequence, the interpretation of crystal lattices in terms of biologically relevant interfaces ceased being a trivial issue and started, in some cases, to be a true challenge. This had to be tackled either by further experimental efforts or by computational means, which required the development of new tools.

A detailed comparison of the features of biologically relevant and of crystal interfaces was described in 1995 (Janin and Rodier, 1995) and followed by a similar study in 1997 (Carugo and Argos, 1997). The two kinds of interfaces were analyzed quantitatively for the first time, by looking at their Buried Surface Area (BSA), defined as the difference in Accessible Surface Area (*ASA*) between uncomplexed and complexed structures. The authors counted both sides of the interface in computing the area, whilst later methods used the average between the two sides, thus dividing this value by 2: $\mathrm{BSA} = 1/2(\mathrm{ASA_u} - \mathrm{ASA_c})$. This latter convention is the one used throughout this review.

The interface classification problem was formulated 2 years later in a seminal paper that also contained the first computational method for interface classification (Janin, 1997). This method relied on a statistical analysis of the interface areas of lattice contacts in crystals of monomeric proteins, which led to an equation relating the BSA to the probability of the interface being a crystal contact.

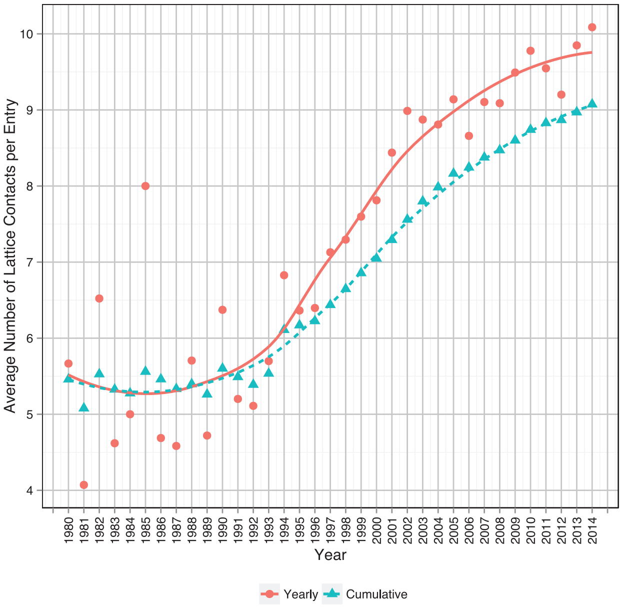

Fig. 2. Average number of contacts between chains of the lattice per PDB entry for structures solved using X-ray crystallography from 1980 to 2014. Averages for all structures solved in a particular year appear in pink (dots, solid line), cumulative averages for the whole PDB per year are shown in turquoise (triangles, dashed line). Essentially the same value for the current average number of contacts was independently obtained in an analysis of a 2012 PDB subset consisting of monomeric proteins with one chain per asymmetric unit (Carugo and Djinović-Carugo, 2012)

In 1998, the Protein Quaternary Structure (PQS) software (Henrick, 1998) provided a quaternary structure estimation for the entire PDB for the first time. To distinguish biological interfaces from crystal contacts, PQS used a composite empirical score based on several geometric and energetic factors: the difference in ASA upon interface formation, the number of buried residues at the interface, an estimation of the difference in the solvation energy of folding of the quaternary assembly and that of its components (Eisenberg and McLachlan, 1986) and the number of salt bridges and interchain disulfide bridges. In 2000, Ponstingl *et al.* studied how to distinguish biological homodimers from crystal dimers by using a knowledge-based atomic pair-potential (Ponstingl *et al.*, 2000). Later, they generalized the method (Ponstingl *et al.*, 2003) to predict full quaternary structure assemblies combining the pairwise method with a graph partitioning algorithm. The corresponding software, named PITA, is still accessible today.

Not long afterward, methods appeared that focused on an essential difference between biologically relevant protein interfaces and crystal contacts: evolutionary conservation. Biological interfaces are the result of evolution and should bear a recognizable signature of selection pressure, while no such pressure acts at crystal contacts to conserve the sequence. A very simple metric of the selection pressure acting on the residues of a given protein is the Shannon entropy (Shannon, 1948) of the position of a multiple alignment of putative homologs of that protein. In 2001, two groups (Valdar and Thornton, 2001b; Elcock and McCammon, 2001) introduced interface classification methods based on evolutionary conservation of interface residues. In their article, Valdar and Thornton (2001a) assessed interface residue conservation in six families of homodimers, calculated the probability that the observed level of conservation had occurred by chance and concluded that that was not the case. In a follow-up work, they extended the analysis to a much larger set of monomeric and homodimeric proteins and studied the usefulness of residue conservation to identify the biological relevance of a protein. They discovered that conservation, combined with interface size, is a powerful predictor of biological relevance. Similarly, Elcock and McCammon (2001) compared the average Shannon entropies of protein interface and surface residues using a simplified amino acid alphabet. In 2005, another group (Guharoy and Chakrabarti, 2005) addressed the issue of possible biases in interface residue versus surface residue entropy comparisons by running a sequence entropy analysis of interface residues only, after subdividing them into 'core' and 'rim' residues based on the presence of fully buried atoms. The rationale of the method is that, for a biological interface, selection pressure should be stronger on the 'hotspot' core residues than on the rim residues, and this difference should show up in an analysis of the average sequence entropies of the two sets. The 'core' and 'rim' concepts are widely agreed to have a central importance in describing the interfaces, though the different authors did not fully agree on their definition. For instance, a residue can be defined as core if it contains at least one fully buried atom upon interface formation (Bahadur *et al.*, 2003; Chakrabarti and Janin, 2002). Alternative definitions of core residue are based on the change in ASA that a residue undergoes upon interface formation (Levy, 2010; Schärer *et al.*, 2010). Rim residues are commonly defined as those interface residues that are not 'core' (e.g. Schärer *et al.*, 2010).

Another important contribution came in 2004, when Bahadur and colleagues introduced the use of packing density, along with residue propensity and hydrophobic interaction scores, for protein interface classification (Prasad Bahadur *et al.*, 2004). A year later the PISA (Protein Interfaces, Surfaces and Assemblies) method was published, an approach based on thermodynamic estimation of interface stability. PISA is currently the most popular method for predicting biological assemblies. Additional details about PISA are described in the next section.

A notable approach, NOXclass, appeared in 2006 and used six interface properties in a support vector machine three-state classifier distinguishing obligate, non-obligate and crystal packing interactions (Zhu *et al.*, 2006). Another method, DiMoVo (Bernauer *et al.*, 2008), relied on Voronoi tessellation to obtain parameters for interface description and a support vector machine to achieve a two-state classification from those parameters. In the same year, the COMP method was introduced, which uses a linear combination of interface complementarity for three features (electrostatic potential, hydrophobicity and shape of the interfaces) (Tsuchiya, 2008) used previously in the PreBI server (Tsuchiya *et al.*, 2006).

In the meantime, the problem had become better known in the community: a review by Kobe *et al.* (2008) provided interesting examples of difficult interface classification cases that ultimately had to be addressed using hybrid methods, combining the crystallographic information with data from other techniques such as site-directed mutagenesis. In another review (Poupon and Janin, 2010), the entire problem was clearly summarized with the statement 'crystallography is surprisingly poor at establishing the quaternary structure'.

Among the more recent methods are IPAC (Mitra and Pal, 2011), using a naive Bayes classifier with ten geometric and physicochemical features, and EPPIC (Evolutionary Protein Protein Interface Classifier) (Duarte *et al.*, 2012), which relies on evolutionary and geometric information. EPPIC remains under active development and is described in more detail in the next section as an example of an evolution-based method.

Another approach (Liu *et al.*, 2014) employs B-factor–related features for interface classification. Here, the rationale is that the average behavior of B-factors in biological interfaces and in crystal contact residues should be different (Carugo and Argos, 1997). Another recent technique uses a method in machine learning, a random forest, to classify protein interfaces based on a variety of interface properties (Luo *et al.*, 2014). A third recent approach, ECR, uses principal component analysis of geometric and energetic (computational alanine mutagenesis) interface features (Sudarshan *et al.*, 2014).

Protein interface classification methods have also been developed for modeling purposes, i.e. to validate quaternary structure in homology modeling templates. To that end, SWISS-MODEL (Biasini *et al.*, 2014) uses a protocol based on sequence conservation, interface hydrophobicity and interface co-occurrence across potential templates

4 Datasets of Biological Interfaces and Crystal Contacts

An important issue affecting all these methods is the availability of reliable datasets of biological interfaces and crystal contacts for method development and benchmarking. For them to be sufficiently reliable, such datasets are nearly always manually curated. Among the most popular are those compiled by Ponstingl *et al.* (2003) and by Prasad Bahadur *et al.* (2004). Recently, two specialized, manually curated datasets that cover only the most difficult-to-classify range of interface areas were introduced: DCbio and DCxtal (Duarte *et al.*, 2012), covering the 800–2000 Å^2 range of BSA. The PiQSi server (Levy, 2007) also provides a large dataset of biological interfaces via a mix of automated (homology inference) and manual validations. In an attempt to create datasets more than one order of magnitude larger, Baskaran *et al.* (2014) introduced two automatically compiled datasets, BioMany and XtalMany, also selected to span a difficult-to-classify area range (500–2000 Å^2).

5 Commonly used methods: two examples

There exist more than a dozen different methods for the computational classification of protein interfaces, some of which were published as proof-of-concept works while others were implemented as user-friendly, publicly available software and even as web servers. This last feature is particularly important since otherwise the methods would remain out of reach for most structural biologists. In Supplementary Table S2, a chronologically ordered list of such methods is shown as well as some fundamental information on each method. Two well-maintained methods using very different and complementary approaches are discussed below in some detail. PISA attempts to estimate the energy of binding, while EPPIC relies on evolution to try to classify the interfaces.

5.1 PISA: protein interfaces, surfaces and assemblies

In 2005, Krissinel and Henrick (2005, 2007) introduced a method for estimating interface stability based on the binding energy of the interface and the entropy change due to complex formation. The estimated interface stability thus dictates whether it should exist in solution (biological interface) or only in the crystalline state (crystal contact). The method goes through a series of approximations and considers BSA, hydrogen bonds, salt bridges and disulfide bonds in order to estimate changes in free energies. For the entropic part, the translational, rotational, vibrational and surface entropy components are estimated using the subunit mass, surface area, symmetry number and inertia moments. After the approximations, several empirical parameters are left to be fitted to training data (the Ponstingl dataset). The energetic estimations are combined with a graph-search algorithm enumerating all the possible assembly combinations present in the crystal. Importantly, interfaces that would assemble infinitely through pure translations or screw axes ('equivalent monomeric units in parallel orientations') are pruned away in the search. This is a fundamental part of the method, since these imposed geometrical and topological constraints are often enough to filter out crystal contacts even before the thermodynamic estimations are brought into play. It also allows contributions from several weak interfaces to be considered in the overall stability of the complex.

The algorithm was implemented in PISA as an online web server, which has become the *de facto* standard for interface classification in the community.

PISA achieves as high as 90% classification accuracy on the training dataset, with the strongest misclassifications being attributed to differences between experimental and physiological conditions. The program generally returns accurate predictions, even for recently solved protein structures, which have become more complex over the past decade. The main difficulties appearing in the predictions seem to be distinction between higher and lower oligomeric states of the same assembly set and the presence of artifactual small molecules and ions in the crystal (Krissinel, 2011). Unlike small molecules and ions, crystallographic water molecules are not taken into account in PISA calculations.

The final quaternary structure predictions are given as a list with the most probable ones appearing first. The sorting is not done only through the free energy values but is helped by a set of rules, preferring for instance higher order oligomers (Krissinel, 2011). The need for these heuristic rules indicates that the approximations leading to the free energy estimations are not accurate enough to be used as the only classification criterion.

Recently, Taudt *et al.* (2015) used a more complex model to provide more accurate estimations of free energies, based on molecular dynamics simulations. The authors compared their results to PISA predictions and found that, although the values were often in discord, the general trends were the same; the rankings of interfaces from most to least stable tended to agree with one another. In the future, it will be interesting to study the agreement of these free energy estimation methods with a statistically significant dataset of experimental free energy values. However, such a dataset must span a broad range of conditions to address the dependency of free energy values on experimental conditions. An early example of a dataset of experimental free energy values is PINT (Kumar, 2006); a more recent one with a focus on mutations, is SKEMPI (Moal and Fernandez-Recio, 2012). A new version of PISA, jsPISA (Krissinel, 2015), has been released within the CCP4 suite (Winn *et al.*, 2011). It features a web graphical user interface and a few improvements that aid in the interpretation of the predictions.

5.2 EPPIC: evolutionary protein–protein interface classifier

EPPIC was published 3 years ago (Duarte *et al.*, 2012), based on earlier work by the same team (Schärer *et al.*, 2010) and is a collection of three different classifiers: one based on geometrical features of the interface and two based on evolutionary features. The evolutionary conservation of residues is assessed by constructing a multiple sequence alignment of all sequence homologs to the target protein structure under study. Differently from previous approaches (Elcock and McCammon, 2001; Glaser *et al.*, 2004; Lichtarge *et al.*, 1996; Valdar and Thornton, 2001a), EPPIC uses closely related homologs only (a 60% sequence identity cutoff when selecting sequences for the alignment). This ensures that homologs share high tertiary and especially quaternary structure similarity (Poupon and Janin, 2010). The information entropy per column of the alignment is then calculated using a reduced amino acid alphabet.

The solvent-ASA of residues is used to classify each interface residue as either 'core' (fully buried upon interface formation) or 'rim' (partially buried upon interface formation). These assignments, together with the conservation values, are used to calculate the two evolutionary scores: core-rim, comparing relative conservation of the interface core residues versus the rim residues; core-surface, comparing the conservation of interface core residues versus the rest of the surface, done with a z-score approach through random sampling of surface residues. Additionally, a simple geometry-based classifier estimates the stability of the interface by counting the number of core residues, defined as 95% buried residues (Schärer *et al.*, 2010). The number of core residues was shown to be a good interface classifier on the DCxtal and DCbio datasets (Duarte *et al.*, 2012).

The three scores are combined to form a consensus call through a simple-majority voting scheme. When EPPIC was originally released in 2012, it was trained using the DCbio and DCxtal datasets that focused on difficult-to-classify interfaces: small biological interfaces and large crystal contacts. EPPIC's prediction accuracy was measured as 89% using the Ponstingl dataset and its predictions agree with PISA 88% of the time on a PDB-wide scale (Baskaran *et al.*, 2014). Unsurprisingly, the lowest agreement between EPPIC and PISA is observed in the 600–1200 Å^2 range of interface area, where classification is particularly hard. The main disadvantage of this evolutionary approach is in situations where not enough sequence data can be found to make a confident prediction. Despite the growth of sequence databases, there are still a certain number of protein structures for which only very few sequence homologs are known. For some cases (especially viral proteins), many sequence homologs are known but with similarity (>90% identity) too high to the studied structures. The alignments resulting from such a distribution of homologs will have little information content compared with varied alignments. These problems in any case will be lessened by the ceaseless growth of sequence databases. In fact, the authors demonstrated that the scores have been improving with the growth of the databases by looking at archived sequence database data from the first 10 years of UniProt (Duarte *et al.*, 2012).

Another possible downside comes in assessing the interfaces between small domains of larger protein structures. EPPIC uses surface residues as the baseline of evolutionary conservation, but in the case of domains the exposed residues in the surface are not necessarily representing the real situation in the full length protein. Thus, scores can be artifactually shifted due to these problems.

6 Comparative methods

In addition to methods based on single structures analysis, several methods exist that incorporate information from multiple structures. Such methods benefit from the redundancy in the PDB and would be expected to gradually increase accuracy and coverage as the PDB grows, in much the same way that homology modeling has benefited from the growth of sequence databases.

6.1 Conservation of interfaces across crystal forms

For many proteins, several structures have been solved via X-ray diffraction performed on different crystal forms: the percentage of PDB entries with at least two crystal forms was estimated to be 64% (Xu and Dunbrack, 2011). For strong biological interfaces, one would expect all crystal forms of the protein to contain the interface, while the crystal contacts might vary depending on the crystal form. Crystallographers have traditionally used this idea to provide evidence for the validity of putative biological interfaces (Gonciarz *et al.*, 2008; Lee *et al.*, 2002). Xu *et al.* (2008) devised a method to perform this analysis automatically, comparing interfaces in different crystal forms across the PDB. They even extended this idea by comparing crystals of homologous proteins. The results were made available via the ProtCID web server (Xu and Dunbrack, 2011). The availability of this resource allowed for some intriguing findings: Weitzner *et al.* (2009) analyzed an unusually small dimer interface in cytosolic sulfotransferases, which was extremely well-conserved across 17 crystal forms, whilst the PDB biological unit annotations were mostly monomeric for these structures. The dimer for the human form of the enzyme was initially identified and thoroughly validated with independent non-crystallographic data (Petrotchenko *et al.*, 2001). In summary, this represents a notable case of a very small, thoroughly validated dimer interface (under 400 Å^2). A limitation of this method resides in the need of several crystal forms for a given interface to robustly assess its biological character, which limits coverage. In addition, there exist cases of crystal contacts conserved across several crystal forms, which represents a source for noise (Xu and Dunbrack, 2011). In an early, detailed study of six crystal forms of the monomeric bovine pancreatic ribonuclease A, it was found that all crystal contacts differed across them with exception of a large dimer found in the three crystal forms where the precipitant was salt (Crosio *et al.*, 1992). Hence, the dimer was interpreted as a salt-induced crystallization intermediate. Such cases suggest that diversity of crystallization conditions across crystal forms of a conserved interface would be an even better criterion for biological relevance.

6.2 Inferring biological units from homology

Close homologs generally have conserved quaternary structure, although some notable exceptions exist (Harrop *et al.*, 2014; Luo *et al.*, 2013; Qin *et al.*, 1998). Thus, if the biological unit of an oligomeric protein or of a protein complex is well-established, one can reasonably assume that proteins with similar sequence retain the same assembly. The key issue is determining a suitable sequence similarity threshold to assume quaternary structure conservation within a protein family. This has been studied by Levy *et al.* (2006, 2008), who analyzed PQS conservation as a function of sequence identity, concluding that 60% identity is a safe threshold for quaternary structure conservation. He also compared quaternary structure information within protein families (Levy, 2007). This analysis resulted in two valuable resources: 3Dcomplex (Levy *et al.*, 2006) and PiQSi (Levy, 2007), which as well as inferring assemblies by homology, allow for community annotation of PDB biological units. An approach, called IBIS, which provides homology-based inference of the most probable biological interactions given a protein structure, was developed in 2010 and uses PISA among other sources of information (Shoemaker *et al.*, 2010). Inferring biological units from homologous templates is of key importance for homology modeling. SWISS-MODEL uses sequence conservation (among other criteria) to assess whether the quaternary structure of a given structural template can be attributed to the sequence to be modeled (Biasini *et al.*, 2014).

7 Assemblies: topology and symmetry

A very important aspect of combining pairwise protein–protein interfaces into stable and finite assemblies is that those assemblies have to fulfill precise topological conditions leading to a small number of closed symmetries. This had already been recognized in the 1960s (Monod *et al.*, 1965) and further elaborated upon in several later works (Hanson, 1966; Levy and Teichmann, 2013). Figure 3 illustrates some basic concepts for homooligomeric assemblies. Homodimer interfaces can be classified into either isologous interfaces, where the same residues participate from both subunits, and heterologous interfaces, where different binding sites are used on each partner. To prevent aggregation and the formation of infinite fibrils, biologically relevant interfaces must either be isologous homodimers (Fig. 3a) or heterologous oligomers with closed symmetry (Fig. 3c and d); other heterologous interfaces lead to infinite assemblies (Fig. 3b).

Higher-order closed assemblies, such as those with dihedral symmetry (Fig. 3e), cannot be formed by heterologous interfaces alone and necessarily also contain isologous interfaces. Such topology and symmetry considerations are very important when it comes to deriving biological assemblies within a given crystal lattice starting from single-interface classification calls. They have also been employed to study how assemblies reflect the evolution of protein complexes and to explain the preponderance of $D(n)$ over $C(n)$ assemblies in the PDB for $n \geq 4$ (Levy *et al.*, 2008).

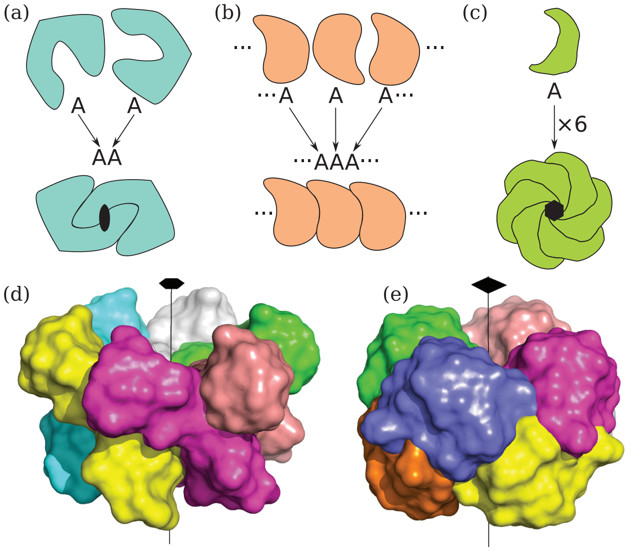

Fig. 3. Protein topology and assemblies. **(a)** Two identical protomers coming together to form an isologous homooligomeric assembly. The black lens denotes a 2-fold axis of rotational symmetry. **(b)** Three identical subunits assembling to form an infinite heterologous homooligomeric assembly. **(c)** Six identical protomers assembling to form a closed heterologous homooligomeric assembly with 6-fold symmetry (the black hexagon). **(d)** A rendering of a cyclic C6 assembly (PDB ID: `4QNB`). **(e)** A protein with dihedral D4 symmetry (PDB ID: `4OAO`)

Importantly, these considerations also apply to heteromers and necessarily lead to even stoichiometry (n:n) as a consequence of the point group symmetry requirement. It is only through pseudosymmetry or self-occlusion that the even stoichiometry rule can be broken. Marsh *et al.* (2015) compiled all known cases of such uneven stoichiometries, finding an occurrence below 10% of all heteromers in the PDB, including some errors in biological unit assignments.

The first method to automatically treat the topology of assembly was PITA (Ponstingl *et al.*, 2003). The authors represented the crystal contacts in a graph with chains as nodes and contacts between them as edges. They additionally weighted the edges with their atomic-pair potential scores. By iterative partitioning of the graph, they were able to enumerate all the closed assemblies and select only those that would score above a certain threshold, finally sorting them by oligomeric size. Such symmetry considerations are also an essential part of the PISA method (Krissinel and Henrick, 2007). With a similar graph representation as PITA, PISA enumerates all possible assemblies in the crystal, rapidly pruning off branches with a combination of closed-symmetry rules and their calculated free-energy scores. Point-group symmetry is also a fundamental part of IPAC, in combination with a naive Bayes classifier (Mitra and Pal, 2011).

In the vast majority of cases, the closed-symmetry requirement is obvious from the crystal packing and automatic methods simply give a quick confirmation of what can be manually observed by the crystallographer. In some cases, however, the assemblies can be very subtle and there automatic methods can help the most. PDB ID `2PEL` (Banerjee *et al.*, 1996), is an example of a subtle lattice arrangement, where partial symmetry exists but there is no closed point-group symmetry beyond a C2 dimer (Supplementary Fig. S1). The annotated biological unit coincides with the tetrameric asymmetric unit; however, the asymmetric unit tetramer has no point group symmetry, so it can only be considered as two copies of a C2 dimer. PISA produces a tetrameric arrangement different from that in the asymmetric unit, but still without a closed point-group symmetry (formed through the largest isologous interface plus a smaller isologous 310 Å^2 interface), showing that the algorithm did not prune off this particular invalid topology. Another interesting case is PDB ID 1R1Z (Velloso *et al.*, 2003), where the authors chose an asymmetric unit with four molecules and no point group symmetry. However, careful observation of the crystal lattice will show that a C4 assembly exists within, which could have some biological relevance (Supplementary Fig. S2). In its predictions, PISA does not show this assembly, which we presume to fall below its scoring threshold.

8 Improving current methods

The variety of approaches employed for protein interface classification makes it particularly difficult to forecast whether the field will see incremental improvements of the existing methods or completely new approaches will appear and become mainstream. Some lines for the future development of the main families of existing methods can; however, be reasonably discerned. For methods based on stability estimations of protein interface stability, new approaches to stability estimation—e.g. those based on advanced force fields and molecular dynamics simulations of the interface of interest (Johnston and Filizola, 2014; Taudt *et al.*, 2015)—may increase accuracy.

For evolution-based methods, various developments can be foreseen. First, the powerful growth of sequence databases will almost certainly continue boosting the coverage and performance of those methods for several years to come. Second, as shown in the past with CRK (Schärer *et al.*, 2010), more sophisticated methods to capture the signal of biological interface evolution than sequence entropy can be employed. A candidate in this area is the correlated mutation approach (Göbel *et al.*, 1994; Pazos *et al.*, 1997; Shindyalov *et al.*, 1994), initially introduced for tertiary structure prediction and subsequently proposed for quaternary structure prediction as well (Hopf *et al.*, 2014; Ovchinnikov *et al.*, 2014). This approach is currently the subject of intensive research, both in terms of its applications in structure prediction and of its foundations and limitations (Talavera *et al.*, 2015). Classifying protein interfaces based on the presence or absence of a subset of contacts exhibiting correlated mutations is an easier task than predicting quaternary assembly *ab initio*, since in the former problem the positioning of the interface partners and the interface geometry are given while in the latter problem only the unbound structures are known. This may help reduce the need for very large multiple sequence alignments that is typical of this approach (Hopf *et al.*, 2014). A challenge for future methods to tackle is the classification of weak biological interfaces, which are particularly difficult since, compared with strong biological interfaces, they are more similar to crystal contacts. The properties of weak interfaces have been studied in detail (Dey *et al.*, 2010; Nooren and Thornton, 2003), which represents a good basis for improving their classification.

9 Interface classification and docking

Protein interface classification has close ties with the field of protein–protein docking. The docking problem can be roughly decomposed into several subproblems (Ehrlich and Wade, 2001; Smith and Sternberg, 2002): (i) sampling: generating the docking poses given the structure of the independent subunits; (ii) scoring: ranking the different poses according to some scoring scheme; and (iii) introducing flexibility in order to refine the good-scoring docking poses. Clearly, the scoring problem is the one that bears close resemblance with interface classification: the sampled protein–protein complexes can be analyzed no differently from any experimentally observed protein–protein interface. Thus, in principle, any of the scoring methods introduced earlier are equally applicable to protein–protein docking.

In fact, the scoring methods developed for the docking field go along the same lines as methods seen earlier: geometrical descriptors (Chen and Weng, 2003; Gabb *et al.*, 1997), energy estimations (Camacho *et al.*, 2000; Norel *et al.*, 2008), knowledge-based statistical potentials (Glaser *et al.*, 2001; Moont *et al.*, 1999; Norel *et al.*, 2008) and evolution-based methods (Choi *et al.*, 2009; Duan, 2005). A set of methods that have also improved docking results are those related to guided or restrained docking (Dominguez *et al.*, 2003). The restraints can come either from experimental data such as chemical shift data or mutagenesis (Dominguez *et al.*, 2003) or from independent predictions to find residues involved in binding (Li and Kihara, 2012; Xue *et al.*, 2014). Surely, there is still potential for cross-breeding between the two fields, which we will hopefully see with new developments in the near future.

10 Conclusion

When compared with other fields and problems in structural bioinformatics, protein interface classification arose later and is a much smaller area of study with very clear practical applications in structural biology. At the same time, it represents a well-defined field with a narrower focus than more classic problems such as tertiary structure prediction, protein–protein docking or protein function prediction. Deeper understanding of the classification problem can help create better foundations for other fields such as protein–protein docking. The number and variety of protein interface classification methods developed in the last decade testifies to the vitality of the field, which appears ready to tackle the challenges of an ever more diverse and complex set of crystal structures and also to find useful applications in other fields for some of its tools and concepts.

Note Added in Proof

With regard to the peanut lectin (PDB ID 2PEL), it is worth noting that in the first paper reporting the structure (Banerjee *et al.*, 1994), the authors described the open tetramer lacking point group symmetry and proposed it as biologically relevant. Later structures of the same protein also contain the open tetramer. Early biochemical studies indicate a pH-dependent dimer–tetramer equilibrium, with a tetramer being the stable form at pH values above 4.75 (Fish *et al.*, 1978). However, this tetramer is not necessarily the same as the open one observed in the crystal structures, and in our opinion the correct assembly should have D2 symmetry. We will carry out further studies to clarify the nature of the peanut lectin tetramer.

Acknowledgement

G.C. thanks Dr Sasa Bjelic and Aleix Lafita for useful discussions.

Funding

Financial support to G.C. from the Swiss National Science Foundation (grant 31003A_140879) and the Research Committee of the Paul Scherrer Institute (grants FK-05.08.1, FK-04.09.1) is gratefully acknowledged, as is IT support from SyBIT/SIS. This research was supported in part by the Intramural

Research Program of the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (support to S.B.).

Conflict of Interest: none declared.

References

Ali,M.H. and Imperiali,B. (2005) Protein oligomerization: how and why. *Bioorg. Med. Chem.*, **13**, 5013–5020.

Astegno,A. *et al.* (2015) Functional roles of the hexamer organization of plant glutamate decarboxylase. *Biochim. Biophys. Acta.*, **1854**, 1229–1237.

Bahadur,R.P. *et al.* (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

Banerjee,R. *et al.* (1994) Crystal structure of peanut lectin, a protein with an unusual quaternary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 227–231.

Banerjee,R. *et al.* (1996) Conformation, protein–carbohydrate interactions and a novel subunit association in the refined structure of peanut lectin-lactose complex. *J. Mol. Biol.*, **259**, 281–296.

Bartesaghi,A. *et al.* (2015) 2.2 Å resolution cryo-EM structure of *β*-galactosidase in complex with a cell-permeant inhibitor. *Science*, **348**, 1147–1151.

Baskaran,K. *et al.* (2014) A PDB-wide, evolution-based assessment of protein–protein interfaces. *BMC Struct. Biol.*, **14**, 22.

Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bernauer,J. *et al.* (2008) DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, **24**, 652–658.

Biasini,M. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–W258.

Camacho,C.J. *et al.* (2000) Scoring docked conformations generated by rigid-body protein–protein docking. *Proteins*, **40**, 525–537.

Campbell,M.G. *et al.* (2015) 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *eLife*, **4**.

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

Carugo,O. and Djinović-Carugo,K. (2012) How many packing contacts are observed in protein crystals? *J. Struct. Biol.*, **180**, 96–100.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–43.

Chen,R. and Weng,Z. (2003) A novel shape complementarity scoring function for protein–protein docking. *Proteins*, **51**, 397–408.

Choi,Y.S. *et al.* (2009) Evolutionary conservation in multiple faces of protein interaction. *Proteins*, **77**, 14–25.

Crosio,M.-P. *et al.* (1992) Crystal packing in six crystal forms of pancreatic ribonuclease. *J. Mol. Biol.*, **228**, 243–251.

Dey,S. *et al.* (2010) The subunit interfaces of weakly associated homodimeric proteins. *J. Mol. Biol.*, **398**, 146–160.

Dominguez,C. *et al.* (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Duan,Y. (2005) Physicochemical and residue conservation calculations to improve the ranking of protein–protein docking solutions. *Protein Sci.*, **14**, 316–328.

Duarte,J.M. *et al.* (2012) Protein interface classification by evolutionary analysis. *BMC Bioinformatics*, **13**, 334.

Ehrlich,L.P. and Wade,R.C. (2001) Protein-protein docking. In: Lipkowitz, K.B. and Boyd, B. D. (eds.) *Reviews in Computational Chemistry*. Wiley-Blackwell, New York, pp. 61–97.

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

Elcock,A.H. and McCammon,J.A. (2001) Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. USA*, **98**, 2990–2994.

Fish,W.W. *et al.* (1978) The macromolecular properties of peanut agglutinin. *Arch. Biochem. Biophys.*, **190**, 693–698.

Gabb,H.A. *et al.* (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**, 106–120.

Glaser,F. *et al.* (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.

Glaser,F. *et al.* (2004) The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, **58**, 610–617.

Göbel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Gonciarz,M.D. *et al.* (2008) Biochemical and structural studies of yeast Vps4 oligomerization. *J. Mol. Biol.*, **384**, 878–895.

Guharoy,M. and Chakrabarti,P. (2005) Conservation and relative importance of residues across protein–protein interfaces. *Proc. Natl. Acad. Sci. USA*, **102**, 15447–15452.

Gut,H. *et al.* (2009) A common structural basis for pH- and Calmodulin-mediated regulation in plant glutamate decarboxylase. *J. Mol. Biol.*, **392**, 334–351.

Hanson,K.R. (1966) Symmetry of protein oligomers formed by isologous association. *J. Mol. Biol.*, **22**, 405–409.

Harrop,S.J. *et al.* (2014) Single-residue insertion switches the quaternary structure and exciton states of cryptophyte light-harvesting proteins. *Proc. Natl. Acad. Sci. USA*, **111**, E2666–E2675.

Henrick,K. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.

Hopf,T.A. *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, **3**, 1–17.

Janin,J. (1997) Specific versus non-specific contacts in protein crystals. *Nat. Struct. Biol.*, **4**, 973–974.

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

Johnston,J.M. and Filizola,M. (2014) Differential stability of the crystallographic interfaces of Mu- and Kappa-opioid receptors. *PLoS One*, **9**, e90694.

Kobe,B. *et al.* (2008) Crystallography and protein–protein interactions: biological interfaces and crystal contacts. *Biochem. Soc. Trans.*, **36**, 1438–1441.

Krissinel,E. (2010) Crystal contacts as nature's docking solutions. *J. Comput. Chem.*, **31**, 133–143.

Krissinel,E. (2011) Macromolecular complexes in crystals and solutions. *Acta. Crystallogr. D Biol. Crystallogr.*, **67**, 376–385.

Krissinel,E. (2015) Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Res.*, **43**, W314–W319.

Krissinel,E. and Henrick,K. (2005) Detection of protein assemblies in crystals. In: Berthold, M.R. (ed.) *Lecture Notes in Computer Science, volume 3695 LNBI of Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pp. 163–174.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.

Kumar,M.D.S. (2006) PINT: protein–protein interactions thermodynamic database. *Nucleic Acids Res.*, **34**, D195–D198.

Lee,W.-H. *et al.* (2002) Comparison of different crystal forms of 3-dehydroquinase from Salmonella typhi and its implication for the enzyme activity. *Acta. Crystallogr. D Biol. Crystallogr.*, **58**, 798–804.

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

Levy,E.D. (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.*, **403**, 660–670.

Levy,E.D. and Teichmann,S.A. (2013) Structural, evolutionary, and assembly principles of protein oligomerization. *Prog. Mol. Biol. Transl. Sci.*, **117**, 25–51.

Levy,E.D. *et al.* (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, e155.

Levy,E.D. *et al.* (2008) Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.

Li,B. and Kihara,D. (2012) Protein docking prediction using predicted protein–protein interface. *BMC Bioinformatics*, **13**, 7.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Liu,Q. *et al.* (2014) Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics*, **15**, S3.

Luo,J. *et al.* (2014) Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants. *Proteins*, **82**, 3090–3100.

Luo,M. *et al.* (2013) Structural determinants of oligomerization of Δ1-pyrroline-5-carboxylate dehydrogenase: identification of a hexamerization hot spot. *J. Mol. Biol.*, **425**, 3106–3120.

Marsh,J.A. *et al.* (2015) Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.*, **6**, 6394.

Mitra,P. and Pal,D. (2011) Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. *Structure*, **19**, 304–312.

Moal,I.H. and Fernandez-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.

Monod,J. *et al.* (1965) On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, **12**, 88–118.

Moont,G. *et al.* (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364–373.

Nogales,E. and Scheres,S.H. (2015) Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol. Cell.*, **58**, 677–689.

Nooren,I.M.A. and Thornton,J.M. (2003) Structural characterisation and functional significance of transient protein–protein interactions. *J. Mol. Biol.*, **325**, 991–1018.

Norel,R. *et al.* (2008) Electrostatic contributions to protein–protein interactions: fast energetic filters for docking and their physical basis. *Protein Sci.*, **10**, 2147–2161.

Ovchinnikov,S. *et al.* (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, **3**, e02030.

Pazos,F. *et al.* (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.*, **271**, 511–523.

Petrotchenko, E.V. *et al.* (2001) The dimerization motif of cytosolic sulfotransferases. *FEBS Lett.*, **490**, 39–43.

Ponstingl,H. *et al.* (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.

Ponstingl,H. *et al.* (2003) Automatic inference of protein quaternary structure from crystals. *J. Appl. Crystallogr.*, **36**, 1116–1122.

Poupon,A. and Janin,J. (2010) Analysis and prediction of protein quaternary structure. *Methods Mol. Biol.*, **609**, 349–364.

Prasad Bahadur,R. *et al.* (2004) A dissection of specific and non-specific protein–protein interfaces. *J. Mol. Biol.*, **336**, 943–955.

Qin,B.Y. *et al.* (1998) Structural basis of the tanford transition of bovine *β*-lactoglobulin. *Biochemistry*, **37**, 14014–14023.

Rose,P.W. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.

Schärer,M.A. *et al.* (2010) CRK: An evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins*, **78**, 2707–2713.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

Shindyalov,I. *et al.* (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng. Des. Sel.*, **7**, 349–358.

Shoemaker, B.A. *et al.* (2010) Inferred Biomolecular Interaction Server–a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.*, **38**, D518–D524.

Smith,G.R. and Sternberg,M.J. (2002) Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, **12**, 28–35.

Stehr,H. *et al.* (2010) PDBWiki: added value through community annotation of the Protein Data Bank. *Database (Oxford)*, **2010**, baq009.

Sudarshan,S. *et al.* (2014) Protein–protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *PLoS One*, **9**, e97115.

Talavera,D. *et al.* (2015) Covariation is a poor measure of molecular coevolution. *Mol. Biol. Evol.*, **32**, 2456–2468.

Taudt,A. *et al.* (2015) Simulation of protein association: kinetic pathways towards crystal contacts. *Phys. Rev. E*, **91**, 033311.

Tsuchiya,Y. (2008) Discrimination between biological interfaces and crystal-packing contacts. *Adv. Appl. Bioinform. Chem.*, **5**, 99.

Tsuchiya,Y. *et al.* (2006) PreBI: prediction of biological interfaces of proteins in crystals. *Nucleic Acids Res.*, **34**, W320–W324.

Valdar,W.S. and Thornton,J.M. (2001a) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.

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

Velloso,L.M. *et al.* (2003) The crystal structure of the carbohydrate-recognition domain of the glycoprotein sorting receptor p58/ERGIC-53 reveals an unpredicted metal-binding site and conformational changes associated with calcium ion binding. *J. Mol. Biol.*, **334**, 845–851.

Weitzner,B. *et al.* (2009) An unusually small dimer interface is observed in all available crystal structures of cytosolic sulfotransferases. *Proteins*, **75**, 289–295.

Westbrook,J.D. and Bourne,P.E. (2000) STAR/mmCIF: An ontology for macromolecular structure. *Bioinformatics*, **16**, 159–168.

Winn,M.D. *et al.* (2011) Overview of the CCP 4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.

Xu,Q. and Dunbrack,R.L. (2011) The protein common interface database (ProtCID)–a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.*, **39**, D761–D770.

Xu,Q. *et al.* (2008) Statistical analysis of interface similarity in crystals of homologous proteins. *J. Mol. Biol.*, **381**, 487–507.

Xue,L.C. *et al.* (2014) DockRank: ranking docked conformations using partner-specific sequence homology-based protein interface prediction. *Proteins*, **82**, 250–267.

Zhu,H. *et al.* (2006) NOXclass: prediction of protein–protein interaction types. *BMC Bioinformatics*, **7**, 27.