Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins

Qifang Xu1,2, Adrian A. Canutescu1, Guoli Wang1, Maxim Shapovalov1, Zoran Obradovic2 and Roland L. Dunbrack Jr1*

1Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA
2Center for Information Science and Technology, Temple University, 1805 North Broad Street, Philadelphia, PA 19122, USA

Received 10 April 2008; received in revised form 30 May 2008; accepted 2 June 2008
Available online 7 June 2008

Many proteins function as homo-oligomers and are regulated via their oligomeric state. For some proteins, the stoichiometry of homo-oligomeric states under various conditions has been studied using gel filtration or analytical ultracentrifugation experiments. The interfaces involved in these assemblies may be identified using cross-linking and mass spectrometry, solution-state NMR, and other experiments. However, for most proteins, the actual interfaces that are involved in oligomerization are inferred from X-ray crystallographic structures using assumptions about interface surface areas and physical properties. Examination of interfaces across different Protein Data Bank (PDB) entries in a protein family reveals several important features. First, similarities in space group, asymmetric unit size, and cell dimensions and angles (within 1%) do not guarantee that two crystals are actually the same crystal form, containing similar relative orientations and interactions within the crystal. Conversely, two crystals in different space groups may be quite similar in terms of all the interfaces within each crystal. Second, NMR structures and an existing benchmark of PDB crystallographic entries consisting of 126 dimers as well as larger structures and 132 monomers were used to determine whether the existence or lack of common interfaces across multiple crystal forms can be used to predict whether a protein is an oligomer or not. Monomeric proteins tend to have common interfaces across only a minority of crystal forms, whereas higher-order structures exhibit common interfaces across a majority of available crystal forms. The data can be used to estimate the probability that an interface is biological if two or more crystal forms are available. Finally, the Protein Interfaces, Surfaces, and Assemblies (PISA) database available from the European Bioinformatics Institute is more consistent in identifying interfaces observed in many crystal forms compared with the PDB and the European Bioinformatics Institute’s Protein Quaternary Server (PQS). The PDB, in particular, is missing highly likely biological interfaces in its biological unit files for about 10% of PDB entries.

© 2008 Elsevier Ltd. All rights reserved.

Edited by M. Sternberg

Keywords: structural bioinformatics; X-ray crystallography; protein-protein interactions

Introduction

Many proteins are oligomeric due to the association of identical subunits under physiological conditions. Homo-oligomerization may be part of allosteric regulation1 or contribute to conformational and thermal stabilities2 and to higher binding affinity with other molecules. Homodimeric proteins have been found to form interactions with a larger number of other proteins compared with monomeric proteins.3 Multi-merization is particularly common in enzymes,
transcription factors, and signal transduction.\textsuperscript{4} The major driving forces for protein multimerization are shape and charge complementarity between the associating subunits brought about by a combination of hydrophobic and polar interactions.\textsuperscript{5,6} Some proteins oligomerize by domain swapping, in which a segment of monomeric protein is replaced by an identical segment from another subunit and vice versa.\textsuperscript{7,8} Many proteins have different predominant oligomeric states under different physiologically relevant conditions, and these states may have important functional differences. Homodimerization may arise in evolution because of stronger tendencies of identical interfaces to self-associate compared with dissimilar interfaces. Heterodimers of proteins in the same superfamily may then evolve from such homodimers.\textsuperscript{9}

Some human diseases are caused by inherited missense mutations in proteins that cause disease in part by having an effect on oligomeric association. For instance, infantile cortical hyperostosis (Caffey disease) is a genetic disorder caused by a missense mutation in exon 41 of the gene encoding the \( \alpha_1(I) \) chain, producing abnormal disulfide-bonded dimeric \( \alpha_1(I) \) chains.\textsuperscript{10} Myofibrillar myopathy is a human disease of muscle weakness, and a causative mutation is localized in the dimerization domain of the filamin C gene, disrupting its secondary structure, leading to an inability to dimerize properly.\textsuperscript{11} Cu/Zn superoxide dismutase (SOD) is an efficient enzyme that catalyzes the conversion of superoxide to oxygen and hydrogen peroxide.\textsuperscript{12} Familial amyotrophic lateral sclerosis or Lou Gehrig’s disease is associated with mutations in Cu/Zn SOD.\textsuperscript{13,14} Some mutations destabilize the SOD dimer, causing abnormal aggregation and may be lethal to cells.

Experimental means for determining the size of an oligomer include analytical ultracentrifugation\textsuperscript{15} and gel filtration.\textsuperscript{16} These methods separate proteins and protein complexes based on their size or mass, from which the oligomerization state may be inferred. However, knowing the size of a protein oligomer does not provide information on the interacting surfaces within an oligomer or the overall structure. Combining separation of oligomers with cross-linking and mass spectrometry can be used to determine protein segments that may be in the binding interfaces between monomers.\textsuperscript{17} Fluorescence resonance energy transfer experiments can be used to identify donor–acceptor pairs of residues that must be near each other in a protein complex to identify which of several dimers in an X-ray crystal structure is likely to be physiologically relevant.\textsuperscript{18} NMR can also be used to determine detailed information on the structure and dynamics of protein oligomers in solution. However, the size of proteins that can be studied easily by NMR is limited.

For most proteins, data on oligomeric association size and, in particular, structure come from X-ray crystallography. For many proteins, the size and actual structure of multimers are controversial or unknown and are based only on what is observed in crystal structures, sometimes even a single crystal structure. Both the Protein Data Bank (PDB)\textsuperscript{19} and the European Bioinformatics Institute\textsuperscript{20,21} provide information on “biological units” or assemblies that are the assumed biologically relevant oligomeric structures found within crystals. The PDB’s biological units are based on what authors of structures themselves believe to be the biologically relevant structure, whereas those of the recently developed Protein Interfaces, Surfaces, and Assemblies (PISA) server\textsuperscript{22} from the European Bioinformatics Institute are based on the analysis of interfaces and predicted stability of complexes observed in single crystal structures. The Protein Quaternary Server (PQS)\textsuperscript{23} contains both manual and automated identifications of biological units (E. Krissinel, personal communication). The PDB and PQS usually have one biological unit size for each PDB entry, whereas the PISA server contains multiple oligomeric structures of different sizes for many PDB entries based on chemical thermodynamic calculations on complex stability. The recently developed Protein Quaternary Structure Investigation (PiQsi) database provides manually annotated sizes of biological units from the literature for PDB entries.\textsuperscript{24}

Many databases and analyses have used PDB and PQS biological units to examine the interfaces between protein domains. For instance, PIBASE\textsuperscript{25} provides a list of structures for a query of two Structural Classification of Proteins (SCOP) superfamilies or family designations\textsuperscript{26} and provides access to coordinates for each pairwise interaction. Interactions in the PIBASE are derived from two sources—the author-approved files provided by the PDB (e.g., pdb1ylv.ent), which generally contain the asymmetric unit of the crystal structure and many nonphysiological interactions, and the hypothetical biological units as proposed by the authors of PQS (e.g., lyv.mmol). The emphasis is on characterizing pairwise interfaces in terms of surface area and polar/nonpolar content. PSIMAP (Protein Structural Interactome map)/PSIBASE (database for PSIMAP)\textsuperscript{27} also performs binary searches for two SCOP-defined domains and finds all structures containing interactions between the query domains. Other databases, such as SNAPPI-DB,\textsuperscript{28} SCOPPI,\textsuperscript{29} and iPfam,\textsuperscript{30} also use the SCOP, PFAM, PDB, and PQS to define atomic interactions among protein domains. Databases of this sort are used for statistical analyses of residue contacts across interfaces to develop methods for predicting or scoring interfaces.\textsuperscript{5,29–33} However, if the data in PDB and PQS are incorrect, these analyses are called into question, both in training data and testing data. Homology modeling based on known multimer structures also depends on accurate multimer structures, and incorrect biological inferences can be made when the assumed quaternary structure of the template is incorrect.

We recently compared the biological units in the PDB and PQS for all crystallographic entries in the PDB and found that they agree on only 83% of entries.\textsuperscript{34} The PDB has a higher tendency than PQS to have biological units that are identical with the asymmetric unit of the same structure, indicating perhaps that many authors may make the unwar-
ranted assumption that the asymmetric and biological units are the same. We also found that the PDB and PQS have inconsistent assignments of biological units for proteins in multiple entries in the PDB that all have the same crystal form. This occurs in the PDB for 12% of entries and in the PQS for about 18% of entries. The PDB's assignments may be more consistent merely because a single research group may solve multiple structures within the same crystal form and assign similar biological units to all of them. When the PDB and PQS agree on the size of a monomer for a single PDB entry, they disagree on the orientation and interface between interacting monomers in less than 2% of cases. The PDB and PQS may have different interfaces across a family of closely related or identical proteins.

A number of studies have attempted to differentiate between biological and crystallization-induced contacts. Ponstingl et al. compiled a set of 96 monomers and 76 homodimers in the PDB by reference to the published literature and compared the ability of buried surface area and pair interaction scores to predict biological contacts in crystals. This data set has subsequently used by others as a benchmark for methods that attempt to determine biological assemblies from single crystals. Bahadur et al. assembled a set of interfaces consisting of 70 homodimeric structures, 122 homodimeric structures, and 188 crystal packing interfaces with surface areas greater than 800 Å² and examined the physical properties of the different interface classes. Shoemaker et al. looked for common interfaces in different crystals of identical and homologous proteins, so-called conserved binding modes, in order to identify likely biologically relevant structures.

In this study, we examined thoroughly the interfaces in crystals of single homologous proteins. We attempted to answer several questions. First, when are two crystals of the same or similar proteins really of the same crystal form and when are they not? We surprisingly found that PDB entries with the same space group, asymmetric unit size, and quite similar unit cell dimensions are occasionally different crystal forms as judged by the interfaces and monomer–monomer orientations that exist within the crystal lattice. Conversely, two crystals in different space groups may be quite similar in terms of all or nearly all the interfaces within each crystal. This occurs (1) when one contains a subset of symmetry operators of the other and a larger asymmetric unit and (2) when one is a small distortion of the other such that the space group is different. This analysis helps sort PDB entries within a family into truly different crystal forms.

Second, we examined the hypothesis used by many crystallographers to infer biological interactions: observation of the same interface in different crystal forms of a protein (or members of the same family) suggests that the interface may be biologically relevant. We compared all interfaces in the available crystal forms in each family and determined those shared by two or more crystal forms. We determined the number of crystal forms with the interface, $M$, compared with the total number of different crystal forms in the same family, $N$. We then evaluated the usefulness of these numbers with prior benchmarks on oligomeric interactions and with NMR structures. When $M$ is greater than 4 or 5, and especially when $M$ is close to or equal to $N$, then the observed interfaces are likely to be part of biologically relevant assemblies. We found 36 families in which all $N$ out of $N$ crystal forms contain a particular interface in which $N \geq 10$. These interfaces are very likely to be physiological. We also found that monomers in a benchmark set composed of both the Ponstingl et al. and Bahadur et al. sets tend to have $M \ll N$.

Third, we examined the usefulness of evolutionary information in evaluating interfaces appearing in more than one crystal form. It occurs often that different crystal forms of identical proteins contain common interfaces but that these usually appear in only two or three such forms and are not shared by homologous proteins. That is, they are only formed under nonphysiological crystallization conditions including high protein concentration, peculiar pH, and the presence of nonphysiological ligands. This has previously been observed for the T4 lysozyme, which has been studied in many crystal forms. When an interface is shared in two crystal forms by divergent proteins, then the interface is very likely to be biologically important. We also found that some interfaces in large families are restricted to one branch of a family, indicating the evolution of an interface in one branch of the family and/or loss in another. This highlights the importance of solving structures of related proteins.

Finally, we compared interfaces common to multiple crystal forms with the annotations found in the PDB, PQS, and PISA server. With an increasing number of crystal forms that contain a given interface, it becomes increasingly likely that the available annotations agree that such an interface is part of a biologically relevant assembly. The PISA server is found to be the most reliable in identifying interfaces for which the evidence, in terms of number of crystal forms containing the interface, seems very high. The PISA server is therefore the best source of biological assembly information when only one or two crystal forms are currently available.

This study is closest to the work of Shoemaker et al., albeit with some important differences. First, we examined the interfaces across PDB entries of homologous proteins to determine whether they are the same crystal form, despite similarities and differences in space group, asymmetric unit size, and unit cell dimensions and angles. Shoemaker et al. separated crystal forms only by space group and/or differences in cell dimensions greater than 2%. We found that these are inadequate to classify crystals as similar or different. Second, we evaluated the usefulness of the number of different crystal forms and the evolutionary relationships of shared interfaces, neither of which was considered by Shoemaker et al. Finally, we provide in the Supplementary Data coordinate files of the shared interfaces that may be useful for further research as training or testing data.
Results

In this study, we are focused on homo-oligomeric structures, hence only PDB entries with a single polypeptide sequence and no nucleic acid present in the crystal. We used SCOP 1.73 to divide 16,164 entries in the PDB containing one protein sequence into families. Because this version of SCOP covers less than 70% of the current PDB, we used PSI-BLAST to assign additional single-sequence entries to SCOP families (see Materials and Methods), for a total of 19,842 entries.

Crystal form redundancy

We are interested in grouping PDB entries in families into different crystal forms in order to test the hypothesis that interfaces common to different crystal forms may be of biological significance. It is well known that some crystals may be analyzed in different space groups with different asymmetric unit sizes, most commonly when one space group contains a subset of the symmetry operators of the other. When looking at different crystals of the same protein or related proteins, we observed many PDB entry pairs in different space groups that, when viewed with molecular graphics, appear to be essentially the same crystal form. As an example, we show in Fig. 1 (top) the crystal lattice of PDB entries 1IJV and 2NLQ, the protein defensin, with different space groups, $P_{21}2_12_1$ and $C_{12}1$, respectively. We therefore expended significant effort in grouping PDB entries into similar crystal forms and groups of crystal forms.

As a first step, we compared PDB entries if they had the same sequence (same length and 100% identity), the same space group, the same asymmetric unit size, and unit cell dimensions and angles within 1% of each other. There was a total of 2991 such entry pairs in our data set with identical sequences. In order to verify that crystals in each pair were truly similar, we compared the interfaces of one entry in each pair with those in the other entry. We used a slightly modified version (see Materials and Methods) of the function $Q$, which we have developed previously to compare interfaces in different pairs. $Q$ expresses the fraction

Fig. 1. Examples of similar crystal structures from different space groups. Top: 1IJV with space group $P_{21}2_12_1$ and asymmetric unit A2 and 2NLQ with space group C121 and asymmetric unit A4 are similar crystal forms. All unique interfaces of one structure are in the other structure with $Q \geq 0.1$ and ASA $\geq 200 \text{ Å}^2$. Bottom: 1ED6 and 1M9J with the same space group $P_{21}2_12_1$, the same asymmetric unit A2, and quite similar unit cell parameters are not the same crystal form. However, all unique interfaces in 1M9J are also in 1ED6, whereas the reverse is not true.
Interface Similarity in Homologous Protein Crystals

Table 1. Numbers of the same crystal structures with various minimum surface areas and minimum Q scores

<table>
<thead>
<tr>
<th>Surface Area (Å²)</th>
<th>500</th>
<th>1000</th>
<th>1500</th>
<th>2000</th>
<th>2500</th>
<th>3000</th>
<th>3500</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.25</td>
<td>0.30</td>
<td>0.35</td>
<td>0.40</td>
<td>0.45</td>
<td>0.50</td>
<td>0.55</td>
</tr>
<tr>
<td>0.05</td>
<td>2567</td>
<td>2567</td>
<td>2567</td>
<td>2567</td>
<td>2567</td>
<td>2567</td>
<td>2567</td>
</tr>
<tr>
<td>0.10</td>
<td>2549</td>
<td>2549</td>
<td>2549</td>
<td>2549</td>
<td>2549</td>
<td>2549</td>
<td>2549</td>
</tr>
<tr>
<td>0.15</td>
<td>2539</td>
<td>2539</td>
<td>2539</td>
<td>2539</td>
<td>2539</td>
<td>2539</td>
<td>2539</td>
</tr>
<tr>
<td>0.20</td>
<td>2523</td>
<td>2523</td>
<td>2523</td>
<td>2523</td>
<td>2523</td>
<td>2523</td>
<td>2523</td>
</tr>
<tr>
<td>0.25</td>
<td>2507</td>
<td>2507</td>
<td>2507</td>
<td>2507</td>
<td>2507</td>
<td>2507</td>
<td>2507</td>
</tr>
<tr>
<td>0.30</td>
<td>2491</td>
<td>2491</td>
<td>2491</td>
<td>2491</td>
<td>2491</td>
<td>2491</td>
<td>2491</td>
</tr>
<tr>
<td>0.35</td>
<td>2474</td>
<td>2474</td>
<td>2474</td>
<td>2474</td>
<td>2474</td>
<td>2474</td>
<td>2474</td>
</tr>
</tbody>
</table>

Table 2. Numbers for structure pairs with identity <100%

<table>
<thead>
<tr>
<th>ID</th>
<th>0</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>350</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>919</td>
<td>919</td>
<td>919</td>
<td>919</td>
<td>919</td>
<td>909</td>
<td>899</td>
<td>862</td>
</tr>
<tr>
<td>0.05</td>
<td>699</td>
<td>839</td>
<td>931</td>
<td>901</td>
<td>907</td>
<td>902</td>
<td>894</td>
<td>856</td>
</tr>
<tr>
<td>0.10</td>
<td>688</td>
<td>834</td>
<td>888</td>
<td>899</td>
<td>905</td>
<td>900</td>
<td>892</td>
<td>856</td>
</tr>
<tr>
<td>0.15</td>
<td>684</td>
<td>831</td>
<td>885</td>
<td>897</td>
<td>903</td>
<td>899</td>
<td>891</td>
<td>855</td>
</tr>
<tr>
<td>0.20</td>
<td>682</td>
<td>826</td>
<td>878</td>
<td>890</td>
<td>898</td>
<td>895</td>
<td>887</td>
<td>853</td>
</tr>
<tr>
<td>0.25</td>
<td>680</td>
<td>823</td>
<td>878</td>
<td>890</td>
<td>896</td>
<td>893</td>
<td>885</td>
<td>851</td>
</tr>
<tr>
<td>0.30</td>
<td>678</td>
<td>822</td>
<td>876</td>
<td>890</td>
<td>896</td>
<td>893</td>
<td>885</td>
<td>850</td>
</tr>
<tr>
<td>0.35</td>
<td>669</td>
<td>813</td>
<td>869</td>
<td>883</td>
<td>888</td>
<td>881</td>
<td>874</td>
<td>847</td>
</tr>
</tbody>
</table>

A structure pair is between a representative entry and an entry with minimum sequence identity in each crystal form. Interfaces with ASA ≥ ASAout are in the other entry and vice versa. Row, surface areas; column, Q scores.

of interacting residues that are similar between two interfaces and is similar to scores used to judge docking success in the CAPRI experiments. A Q value of 1 means that all contacts in one interface are shared in the other interface and vice versa. A Q value of 0 means that no contact is shared. Random simulations indicate that a Q value ≥ 0.1 indicates a statistically significant similarity of two interfaces (see Materials and Methods).

We compared crystals using different minimum values of Q and minimum values of the accessible surface area (ASA) of unique interfaces that exist in each crystal. For the first member of each pair (entry A), we took all interfaces with a surface area greater than or equal to ASAmin and compared each interface with all the interfaces in the second member of the pair (entry B), regardless of the surface area in entry B. The reason for this is that an interface with an area just over ASAmin in entry A might correspond to an interface in entry B with an area just below ASAmin, which would be missed if ASAmin was applied to entry B. We calculated fAB as the fraction of interfaces in entry A with a surface area greater than or equal to ASAmin that have interfaces in entry B with Q ≥ Qmin. If fAB = 1, then all interfaces in entry A have corresponding interfaces in entry B given the ASAmin and Qmin cutoffs. This process was then performed in reverse (interfaces of entry B with ASA ≥ ASAmin against all interfaces in entry A) to calculate fBA given the same cutoffs.

In Table 1, we show the number of same-sequence entry pairs for which fAB = 1 and fBA = 1 for various values of ASAmin and Qmin. First, for Qmin = 0 and larger minimum surface areas, the number drops below 2991, because a small number of crystals have no interface larger than those values of ASAmin; therefore, fAB and fBA are undefined. For Qmin values >0, the number of pairs with fAB = 1 and fBA = 1 drops off at lower values of ASAmin. We examined visually the seven entry pairs that do not have fAB = 1 and fBA = 1 with Qmin = 0.1 and ASAmin = 100 Å². We found four entry pairs in this category that are actually different crystal forms as judged by the packing of monomers within the crystal. These pairs all contained the same sequence of rainbow trout lysozyme: PDB entries 1LMN, 1LMC, 1BB6, and 1BB7. Although 1LMN and 1LMC are visually the same crystal as are 1BB6 and 1BB7, 1LMN and 1LMC are different from 1BB6 and 1BB7. The other three pairs are similar in all respects, except for small interfaces containing loops that are present in one crystal but absent or perturbed in the other.

We performed the same analysis on different-sequence pairs. For this analysis, we first removed redundant entries from the set considered in Table 1. That is, we only used one entry for each sequence among those entries with similar crystal parameters. We split the rainbow trout lysozyme entries into two crystal forms (1LMN and 1LMC on one hand, and 1BB6 and 1BB7 on the other). This resulted in 919 entry pairs for comparison using various cutoffs of ASAmin and Qmin as shown in Table 2. This time, for ASAmin ≥ 250, some entry pairs have no interface above ASAmin, and so fAB and fBA are undefined. We examined visually the 31 entry pairs for which either fAB or fBA = 1 was ≤ 1 when ASAmin = 100 and Qmin = 0.1. Thirty of these pairs appear to be the same crystal forms, with very similar lattice structures but with differences in some small interfaces due to loop movements. One pair of the 31 pairs, which was composed of PDB entries 1DPO (anionic trypsin) and 1ZSL (factor Xa) (39% sequence identity), have different crystals. These entry pairs are of the same space group (I23), have the same asymmetric unit size (monomer), and have the same unit cell constants that differ by ≤ 1%, and yet the crystals are different. This indicates that some care should be shown to verify that entries in the PDB that appear to be similar crystal forms really are the same.

At a Qmin value of 0.1, the highest detection rate of similar crystals (i.e., with both fAB = 1 and fBA = 1) occurs for ASAmin = 200 Å² (in fact, at all Qmin values >0, this is true). If we define false negatives as crystals that are actually the same crystals but fail to meet these criteria, with the five crystal entry pairs described above (true negatives), these cutoffs contain only 14 of 3910 false negatives (2991 + 919 – 2898 – 905 – 5 = 14), or 0.36%. False positives are crystal pairs that satisfy the criteria but are actually different. After checking visually dozens of entry pairs with both fAB = 1 and fBA = 1, we found no false positive, but it is difficult to calculate a true false-positive rate. Given the same space group, the same asymmetric unit, and tight constraints on the unit cell dimensions, it is likely to be very low.
Using the representatives described above, we then compared entries in each family with different space groups, asymmetric unit sizes, and/or unit cell dimensions. Using cutoffs of when \( \text{ASA}_{\text{min}} = 200 \) and \( Q_{\text{min}} = 0.1 \), we found 6931 pairs (of 276,643) with \( f_{AB} = 1 \) and \( f_{BA} = 1 \). Most of these pairs (90%) had high sequence identity (\( \geq 90\% \)). There were 1375 pairs in different space groups.

We decided to examine additional pairs for which \( f_{AB} = 1 \) or \( f_{BA} = 1 \) but not both. In these cases, all the interfaces in one crystal with \( \text{ASA} \geq 200 \text{ Å}^2 \) are contained in the other crystal, but the reverse is not true. We examined 100 of these visually to determine whether the crystals were the same or different. In most cases, the crystals appeared to be the same, preserving the orientations and interactions of all proteins within the crystal. In some cases, the crystals were visually different but contained some obvious similarities that are not likely to be biological interactions. This occurs, for instance, when planes of proteins are similar in the two crystals but adjacent planes pack differently, either shifted relative to each other or in a back-to-front fashion in one crystal while in a back-to-back fashion in the other. An example of this is shown in Fig. 1 (bottom). The figure shows crystals of bovine and human nitric oxide synthase, PDB entries 1ED6 \(^{43}\) and 1M9J \(^{44}\), respectively. Horizontal planes of proteins in the two crystals (as oriented in the figure) are quite similar, but the packing modes of these planes are different. If the top and bottom horizontal planes are rotated 180° about a vertical axis and shifted, then one crystal can be obtained from the other.

Since we would like to group entries that contain similar nonbiological interfaces, those that would produce infinite lattices, together, we decided to group together entries that have \( f_{AB} = 1 \) and \( f_{BA} \geq 0.65 \) or \( f_{AB} \geq 0.65 \) and \( f_{BA} = 1 \). These occur for a further 2883 entry pairs for the representatives described above. The value of 0.65 was chosen after inspecting many entry pairs. If we group entries that share \( f_{AB} = 1 \) and \( f_{BA} \geq 0.65 \) or \( f_{AB} \geq 0.65 \) and \( f_{BA} = 1 \) with complete linkage clustering, the result is a total of 8816 groups. This is down from a total of 12,394 crystal forms if space groups, asymmetric unit sizes, and unit cell parameters (within 1%) are required to be the same and \( f_{AB} = 1 \) and \( f_{BA} = 1 \). We call the 8816 groups \textit{crystal form groups} or CFGs since they may contain crystals that contain different orientations of monomers although they share most or all crystal-induced interactions. The full procedure for defining these CFGs is shown in Fig. 2.

### Common interfaces in multiple CFGs

Some statistical data on these CFGs are shown in Table 3. There are 1125 SCOP-defined families with at least two CFGs. Most CFGs contain highly similar sequences, with only 5% of CFGs with a minimum sequence identity <90% (Fig. 3, top left), whereas between any two CFGs in a family, 95% of CFG pairs have a minimum pairwise sequence identity <90% (Fig. 3, top right). The number of CFGs in a family ranges from 1 to 173 (SCOP family protein kinases, catalytic subunit, d.144.1.7), and 51% of families have two or three CFGs (Fig. 3, bottom left).

We define the interfaces contained in a single CFG as those that exist in all entries in the group. The interfaces of the entry with the best resolution were compared with interfaces from all other entries in the CFG; the interfaces in the CFG were defined as those in the representative entry with \( Q \geq 0.1 \) to at least one interface in each of the other members of the CFG. We then compared interfaces in different CFGs using a \( Q_{\text{min}} \) value of 0.1. Table 1 also gives an overview of the total common interfaces. In total, 868 families have at least one common interface in two or more CFGs, involving 15,264 entries and 3771 common interfaces (some CFGs have multiple common interfaces; e.g., if they share tetramers). Of
these, 579 families have all CFGs (N out of \(N\) CFGs in the family) sharing at least one common interface (Fig. 3, bottom right). There are 176 families with at least one common interface in all CFGs with \(N \geq 4\), which are composed of 1372 CFGs and 3139 entries. There are 248 families containing 2136 CFGs and 5200 entries that have at least one common interface existing in at least \(M=4\) CFGs, when \(M/N \geq 0.5\) (\(N\) is the total number of CFGs in the family). We examined whether these interfaces existed in the biological unit assemblies as defined in the PDB, PQS, and PISA server. The percentages of common interfaces available in PDB, PQS, and PISA server biological units are also shown in Table 3. For those interfaces in a large number of crystal forms, the PISA server identifies 97% of these interfaces as part of a biological assembly when \(M=4\) and \(N \geq 4\) and 93% when \(M<N, M/N>0.7,\) and \(N \geq 4\).

**Benchmarking the values of \(M\) and \(N\) as indicators of biologically relevant interfaces**

Ponstingl et al.\(^35\) and Bahadur et al.\(^36,37\) established benchmarks of monomers and dimers/multimers as described in the literature in solution experiments. These have been used by others to test methods that distinguish biological from purely crystal-induced interfaces. We combined these sets to form a single benchmark, consisting of 132 monomers, 84 dimers, 15 trimers, 19 tetramers, and 8 hexamers. We first checked these PDB entries to see what their biological unit sizes were in the

---

**Table 3. Overview of common interfaces**

<table>
<thead>
<tr>
<th>Family Type</th>
<th>No. of families</th>
<th>No. of CFGs</th>
<th>No. of entries</th>
<th>In PDB (%)</th>
<th>In PQS (%)</th>
<th>In PISA server (%)</th>
<th>In asymmetric unit (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-domain structures</td>
<td>1377</td>
<td>8816</td>
<td>19842</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CFGs/family ≥ 2</td>
<td>1125</td>
<td>8564</td>
<td>19446</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CFGs/cluster ≥ 2</td>
<td>868</td>
<td>6292</td>
<td>15264</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(N/N (N \geq 4))</td>
<td>176</td>
<td>1372</td>
<td>3139</td>
<td>90</td>
<td>95</td>
<td>97</td>
<td>58</td>
</tr>
<tr>
<td>(1&gt;M/N&gt;0.7 (N \geq 4))</td>
<td>74</td>
<td>676</td>
<td>1781</td>
<td>88</td>
<td>94</td>
<td>93</td>
<td>61</td>
</tr>
<tr>
<td>(N/N (N=2, 3))</td>
<td>266</td>
<td>618</td>
<td>1049</td>
<td>70</td>
<td>76</td>
<td>80</td>
<td>52</td>
</tr>
<tr>
<td>(M \geq 4 (M/N \geq 0.5)^b)</td>
<td>248</td>
<td>2136</td>
<td>5200</td>
<td>89</td>
<td>93</td>
<td>95</td>
<td>59</td>
</tr>
</tbody>
</table>

\(a\) Percentages of entries in the PDB, PQS, PISA, and ASU are based on clusters and may have duplicate entries. NMR entries are excluded.  
\(b\) From the analysis of the Ponstingl/Bahadur benchmark set and NMR structures, \(M \geq 4\) and \(M/N \geq 0.5\) are most likely to be of biological interest.

---

**Fig. 3.** Overview of families and crystal forms in the database. Top left: Sequence identity distribution in the same CFGs. Top right: Sequence identity distribution between different CFGs. Bottom left: Number of families (y-axis) with \(N\) CFGs (x-axis). Bottom right: The relationship between the number of families (y-axis) and the number of CFGs (x-axis) such that all \(N\) CFGs for the family contain at least one common interface. The y-axis is the cumulative number of families such that \(N \geq x\).
existing databases of the PDB, PQS, and PISA server. These results are shown in Table 4. Although the dimers and larger oligomers are mostly correct in the three public databases, the monomers are classified as larger structures one-third of the time by the PDB and PISA server and more than half the time by the PQS. This is mostly due to the monomers in the Bahadur et al. set, which were

Table 4. Biological units of benchmark entries in public databases

<table>
<thead>
<tr>
<th></th>
<th>PDB</th>
<th>PQS</th>
<th>PISA server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark monomers (132)</td>
<td>88 monomers</td>
<td>60 monomers</td>
<td>87 monomers</td>
</tr>
<tr>
<td></td>
<td>41 dimers</td>
<td>55 dimers</td>
<td>36 dimers</td>
</tr>
<tr>
<td></td>
<td>3 larger oligomers</td>
<td>17 larger oligomers</td>
<td>9 larger oligomers</td>
</tr>
<tr>
<td>Benchmark dimers (84)</td>
<td>3 monomers</td>
<td>1 monomer</td>
<td>1 monomer</td>
</tr>
<tr>
<td></td>
<td>78 dimers</td>
<td>76 dimers</td>
<td>80 dimers</td>
</tr>
<tr>
<td></td>
<td>3 larger oligomers</td>
<td>7 larger oligomers</td>
<td>3 larger oligomers</td>
</tr>
<tr>
<td>Benchmark oligomers (size &gt; 2) (42)</td>
<td>1 monomer</td>
<td>1 monomer</td>
<td>1 monomer</td>
</tr>
<tr>
<td></td>
<td>1 dimer</td>
<td>0 dimer</td>
<td>1 dimer</td>
</tr>
<tr>
<td></td>
<td>40 larger oligomers</td>
<td>41 larger oligomers</td>
<td>40 larger oligomers</td>
</tr>
</tbody>
</table>

Fig. 4. Correlations between M and N for the benchmark data set: (a) M versus N for 132 monomers. (b) M versus N for 126 dimers and oligomers. (c) M versus N for interfaces not present in the available NMR structures. (d) M versus N for interfaces that are found in available NMR structures.
chosen to contain crystallographic interfaces larger than 800 Å².

We first examined the monomers in the benchmark set and determined whether any of the interfaces in each crystal appeared in other CFGs in the family. For each entry, we determined \( N \), the number of CFGs in the family of the entry, and then \( M \), the largest number of CFGs that any interface in the benchmark entry shows in accordance to a similarity criterion of \( Q \geq 0.1 \). A bar chart of \( M \) versus \( N \) for the 132 monomers is shown in Fig. 4a. Only 4 of the entries are in families with only 1 CFG. Of the remaining 128 entries with 2 or more CFGs, 55 do not contain interfaces in common with any other crystal form. A further 53 entries contain interfaces common to only 2 or 3 CFGs. Only 12 entries contain interfaces observed in 5 or more CFGs, and these are all in families with 19 or more CFGs, most with greater than 30, as shown in Table 5. The 2 entries in the benchmark with \( M = 14 \) are both a T4 lysozyme, and the associated interface is one previously noted as occurring in many crystal forms by Faber and Matthews.39 A structure of α-bungarotoxin, PDB entry 2abx,45 has a common interface with \( M = 9 \). This interface also exists in the NMR structure of neuronal bungarotoxin (45% identity), PDB entry 2NB, although this structure was solved partly by homology to crystal structures since there were few interprotein nuclear Overhauser enhancements.46 The PDB and PQS have the same dimer, but the dimer from the PISA server contains a different interface.

We examined the 126 dimers and higher oligomers in the benchmark set and determined whether these crystals contain interfaces observed in multiple CFGs. A plot of \( M \) versus \( N \) for the largest \( M \) for each entry is shown in Fig. 4b. The distribution is quite different from that of the monomers. In total, 55 (44%) of the benchmark oligomers contain at least one interface that is present in all \( N \)'s of N CFGs in each family. Seventy-seven of the entries contain interfaces in \( M \geq 5 \) or more CFGs (of 98 available with \( N \geq 5 \), or 79%). Since the benchmark data do not identify which interfaces are biological and only the size of the assembly in solution experiments, we cannot be sure that all these interfaces are biological. Nevertheless, the monomer and oligomer benchmark data shown in Fig. 4a and b and Table 5 indicate than when \( M \geq 5 \) and \( M/N \geq 0.5 \), an entry is likely to contain a biological assembly larger than a monomer.

The Ponstingl et al. set does not contain coordinates for interfaces involved in the dimers and larger oligomers. So we examined the values of \( M \) and \( N \) for the interfaces in two or more CFGs and then determined whether these interfaces were in any available NMR structures for the family. The results are shown in Fig. 4c and d and divided into those interfaces not found in NMR structures (Fig. 4c) and those that are found in at least one NMR structure (Fig. 4d). The figures are quite similar to the Ponstingl et al. set.

When an interface is not in any NMR structure, \( M \) tends to be quite low and \( M \ll N \); that is, there are few points near the diagonal at larger \( N \). We examined the eight cases in Fig. 4c for which \( M/N \geq 0.5 \) and \( N \geq 5 \). In most of these cases, monomeric proteins were specifically chosen for study (data not shown), due to the size limitation of most NMR experiments. When an interface is in an NMR structure, \( M \) is larger (if \( N \) is) and \( M/N \geq 0.5 \), so that most data points are near the diagonal. A table with all common interfaces confirmed by NMR structures is provided in the Supplementary Data.

In order to derive a probability that an interface is biologically relevant of \( M \) and \( N \), we combined the benchmark data set described above and the PIQSi database.22 The PIQSi database is a manually curated set of biological unit sizes for 15,000 entries in the PDB. These were derived from literature references and comparisons across PDB entries within families. The PIQSi database may contain some errors but seems to be carefully curated and so provides a large set of biological unit sizes sufficient for estimating \( P \). The final data set contains 4171 monomers and 3090 dimers/oligomers.

### Table 5. Interfaces with \( M \geq 5 \) from benchmark monomers

<table>
<thead>
<tr>
<th>PDB entry</th>
<th>Family</th>
<th>Cluster</th>
<th>No. entries in family</th>
<th>No. entries in cluster</th>
<th>ASA</th>
<th>Minimum sequence identity</th>
<th>In PDB</th>
<th>In PQS</th>
<th>In PISA server</th>
<th>No. of entries in PDB</th>
<th>No. of entries in PQS</th>
<th>No. of entries in PISA server</th>
</tr>
</thead>
<tbody>
<tr>
<td>232I</td>
<td>34</td>
<td>14</td>
<td>443</td>
<td>413</td>
<td>731</td>
<td>77</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>324</td>
<td>11</td>
</tr>
<tr>
<td>256I</td>
<td>34</td>
<td>14</td>
<td>443</td>
<td>413</td>
<td>762</td>
<td>77</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>324</td>
<td>11</td>
</tr>
<tr>
<td>1caq</td>
<td>38</td>
<td>12</td>
<td>78</td>
<td>31</td>
<td>672</td>
<td>58</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>830I</td>
<td>38</td>
<td>12</td>
<td>78</td>
<td>31</td>
<td>707</td>
<td>58</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>1kwa</td>
<td>44</td>
<td>10</td>
<td>52</td>
<td>12</td>
<td>579</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>1afrk</td>
<td>48</td>
<td>9</td>
<td>175</td>
<td>72</td>
<td>822</td>
<td>63</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>2abx</td>
<td>19</td>
<td>9</td>
<td>31</td>
<td>13</td>
<td>363</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>1clu</td>
<td>137</td>
<td>7</td>
<td>212</td>
<td>34</td>
<td>754</td>
<td>26</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>18</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>1vlz</td>
<td>50</td>
<td>7</td>
<td>94</td>
<td>25</td>
<td>404</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1r3k</td>
<td>32</td>
<td>6</td>
<td>118</td>
<td>25</td>
<td>616</td>
<td>31</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>18</td>
<td>35</td>
<td>21</td>
</tr>
<tr>
<td>1mwc</td>
<td>55</td>
<td>6</td>
<td>337</td>
<td>24</td>
<td>299</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1bin</td>
<td>55</td>
<td>5</td>
<td>337</td>
<td>11</td>
<td>561</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

* Number of entries that contain the common interfaces.
oligomers. Plots similar to Fig. 4 for monomers and dimers/oligomers in the PiQSi database look very similar to Fig. 4a and b, respectively, albeit with somewhat more noise (data not shown). Furthermore, we divided the data set into two sets based on the minimum sequence identity among \( M \) CFGs that contain a common interface: a data set with identity \( \geq 90\% \) (2185 monomers and 254 dimers/oligomers) and a data set with identity \( < 90\% \) (1986 monomers and 2836 dimers/oligomers). If an interface exists in one CFG only, the minimum identity is set to be 100%. The probability \( P(\text{biological interface}|M,N) \) is calculated by

\[
P = \frac{\text{Number of Oligomers}(M,N)}{\text{Number of Monomers}(M,N) + \text{Number of Oligomers}(M,N)}
\]

for each \( M,N \) combination, where the number of oligomers includes both dimers and larger oligomeric structures.

Table 6 shows the probabilities for each \( M \) and \( N \) from the data set with identity \( \geq 90\% \). The probabilities for each \( M \) and \( N \) from the data set with identity \( < 90\% \) are listed in Table 7. Tables containing the numbers of entries in each cell in Tables 6 and 7 are provided in the Supplementary Data. When proteins share high sequence identity (>90%), there is some tendency for similar interfaces to be used in different crystal forms even in cases in which the proteins are likely to be monomers under approximately physiological conditions. However, when there is some sequence divergence (\( \leq 90\% \) identity), \( P \geq 0.95 \) even when \( M \geq 2 \) and \( M \) is close to \( N \). The results emphasize the value of having crystal structures of more than one member per family.

**Interfaces present in large numbers of CFGs**

We examined families for which all available CFGs contain one or more common interfaces (Fig. 3, bottom right). We found 36 families (594 crystal forms, 1585 entries) with at least one common interface in all \( N \) crystal forms where \( N \geq 10 \) and 18 families when \( N \geq 15 \). Table 8 presents these 18 families. A table with \( N \geq 2 \) is presented in the Supplementary Data. The interfaces described in Table 6 are almost certainly biological interactions, since it is highly unlikely that a nonbiological interface could form under 15 or more crystallization conditions. These are well-studied proteins, and the interfaces are commonly known: the PDB biological units contain the common interfaces in 90% of these entries, whereas the PiQSi biological units contain the interfaces in 94% of the entries. For the PISA server, we take the most stable assembly defined for each PDB entry; the PISA server contains these common interfaces for 96% of these entries. Only 62% of these entries contain the common interfaces in their asymmetric units. The interfaces range in size from 900 to 4800 Å\(^2\). Some of the families contain more than

**Table 6. Probability \([P(\text{biological interface}|M,N)]\) for each \( M, N \) and minimum identity \( \geq 90\% \), derived from 2185 monomers and 254 dimers/oligomers in the benchmark data set and the PiQSi database**

<table>
<thead>
<tr>
<th>( M )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>&gt;14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.15</td>
<td>0.13</td>
<td>0.18</td>
<td>0.07</td>
<td>0.04</td>
<td>0.24</td>
<td>0.08</td>
<td>0.06</td>
<td>0.09</td>
<td>0.50</td>
<td>0.50</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.58</td>
<td>0.40</td>
<td>0.09</td>
<td>0.07</td>
<td>0.03</td>
<td>0.18</td>
<td>0.07</td>
<td>0.13</td>
<td>0.27</td>
<td>0.27</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
<td>0.25</td>
<td>0.60</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&gt;14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Row, \( M \); column, \( N \). Empty cells mean that there is no data point available. Values in boldface are total data points \( \leq 10 \).

**Table 7. Probability \([P(\text{biological interface}|M,N)]\) for each \( M, N \) and minimum identity \( < 90\% \), derived from 1986 monomers and 2836 dimers/oligomers in the benchmark data set and the PiQSi database**

<table>
<thead>
<tr>
<th>( M )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>&gt;14</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.95</td>
<td>0.08</td>
<td>0.3</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.15</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>0.10</td>
<td>0.67</td>
<td>0.76</td>
<td>0.88</td>
<td>0.1</td>
<td>0.31</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;14</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.67</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Excluding 358 lysozyme entries with 14/34 common interfaces and minimum identity = 77.
one shared interface, implying the existence of common tetramers and larger structures. For instance, the l-aspertase/fumarase family (SCOP a.127.1.1) has three common interfaces.

Some SCOP families are broader than others, containing relatively distant homologues and clear paralogues. At larger evolutionary distances, we expect that some biologically relevant interfaces will not be conserved or will have evolved in one branch of a family but not in others.47 It is always possible that some crystals form under conditions in which dimers and larger structures do not form or in which only some monomers enter the crystal.18 We therefore examined families for which M of N crystal forms contained an interface, where M was less than N but M was still fairly large. Table 9 gives 16 families (511 crystal forms, 1437 entries) with common interfaces in 4 crystal forms (7 entries) with a surface area of 2722.2 Å². These 7 entries are remotely related to the entries in cluster 1 with an identity less than 10%. The PDB, PQS, and PISA server all give the same biological dimers with the common interface. The entry 1OTK, phenylacetic acid degradation protein, is distantly related to all other entries in this family with a sequence identity less than 10%. The PDB, PQS, and PISA server all give the same biological dimers with the common interface.

We further analyzed the tendency for PDB biological units to contain interfaces found in multiple CFGs (Fig. 6). The larger the value of M is, the more likely that the PDB, PQS, and PISA server biological units will contain an interface when

Table 8. Interfaces in N of N crystal forms (N ≥ 15) in a SCOP-defined family

<table>
<thead>
<tr>
<th>Family name</th>
<th>SCOP code</th>
<th>Family N</th>
<th>No. of entries in family</th>
<th>PDB biological unit</th>
<th>PQS biological unit</th>
<th>PISA server biological unit</th>
<th>Asymmetric unit</th>
<th>ASA</th>
<th>Minimum identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAT-like</td>
<td>c.67.1.1</td>
<td>42</td>
<td>152</td>
<td>149</td>
<td>149</td>
<td>152</td>
<td>83</td>
<td>2998</td>
<td>7</td>
</tr>
<tr>
<td>GABA aminotransferase-like</td>
<td>c.67.1.4</td>
<td>42</td>
<td>123</td>
<td>104</td>
<td>119</td>
<td>123</td>
<td>74</td>
<td>3791</td>
<td>6</td>
</tr>
<tr>
<td>Cystathionine synthase-like</td>
<td>c.67.1.3</td>
<td>32</td>
<td>51</td>
<td>47</td>
<td>50</td>
<td>51</td>
<td>33</td>
<td>2560</td>
<td>6</td>
</tr>
<tr>
<td>Crotonase-like</td>
<td>c.14.1.3</td>
<td>26</td>
<td>35</td>
<td>31</td>
<td>34</td>
<td>34</td>
<td>26</td>
<td>1459</td>
<td>12</td>
</tr>
<tr>
<td>PNP oxidase-like</td>
<td>b.45.1.1</td>
<td>24</td>
<td>36</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>20</td>
<td>1761</td>
<td>6</td>
</tr>
<tr>
<td>Decarboxylase</td>
<td>c.1.2.3</td>
<td>22</td>
<td>55</td>
<td>42</td>
<td>55</td>
<td>54</td>
<td>35</td>
<td>1815</td>
<td>12</td>
</tr>
<tr>
<td>TNF-like</td>
<td>b.22.1.1</td>
<td>20</td>
<td>28</td>
<td>27</td>
<td>27</td>
<td>28</td>
<td>20</td>
<td>982</td>
<td>5</td>
</tr>
<tr>
<td>Nucleoside diphosphatase kinase</td>
<td>d.58.6.1</td>
<td>19</td>
<td>59</td>
<td>59</td>
<td>58</td>
<td>58</td>
<td>33</td>
<td>893</td>
<td>25</td>
</tr>
</tbody>
</table>

Fig. 5, those entries that contain this interface are individually boxed. Entries are colored by their CFG. The PDB has the common interface in all its biological units. The PQS has three monomers (2O1Z, 1SYY, and 2ANJ), and the PISA server has one monomer (1SMQ) and one dimer (2P1I) that is not the common dimer in the 11 crystal forms. The other common interface (Fig. 5, inset) is observed in the steeroyl-acyl carrier desaturases in 4 crystal forms (7 entries) with a surface area of 2722.2 Å². These 7 entries are remotely related to the entries in cluster 1 with an identity less than 10%. The PDB, PQS, and PISA server all give the same biological dimers with the common interface.

For many of these families, phylogenetic trees show that the interface is conserved within an entire branch of the family. We give the ribonucleotide reductase-like family (SCOP a.25.1.2) as an example. Members of this family have been demonstrated to be homodimers48–50 in gel electrophoresis experiments. A phylogenetic tree of the PDB entries in this family is shown in Fig. 5. An analysis of the common interfaces in different crystal forms shows that there are two types of dimers. One is a common interface, cluster 1, that occurs in 11 crystal forms (41 entries) with an average ASA of 1985.3 Å². In the main part of Fig. 5, those entries that contain this interface are

† http://dunbrack.fccc.edu/JMB08/Supplementary Figures.tar
Table 9. SCOP families with $N \geq 16$, $M \geq 7$, and $M/N \geq 0.7$

<table>
<thead>
<tr>
<th>Family name</th>
<th>SCOP code</th>
<th>Family $N$</th>
<th>Cluster $M$</th>
<th>No. of entries in family</th>
<th>No. of entries with interface</th>
<th>PDB biological unit</th>
<th>PQS biological unit</th>
<th>PISA server biological unit</th>
<th>Asymmetric unit</th>
<th>ASA</th>
<th>Minimum identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine-dependent oxidoreductases</td>
<td>c.2.1.2</td>
<td>147</td>
<td>111</td>
<td>338</td>
<td>251</td>
<td>233</td>
<td>247</td>
<td>247</td>
<td>148</td>
<td>1693</td>
<td>4</td>
</tr>
<tr>
<td>Triosephosphate isomerase</td>
<td>c.1.1.1</td>
<td>39</td>
<td>33</td>
<td>81</td>
<td>75</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>65</td>
<td>1587</td>
<td>17</td>
</tr>
<tr>
<td>Tryptophan synthase beta subunit-like PLP-dependent enzymes</td>
<td>c.79.1.1</td>
<td>34</td>
<td>26</td>
<td>48</td>
<td>39</td>
<td>35</td>
<td>39</td>
<td>38</td>
<td>31</td>
<td>2099</td>
<td>11</td>
</tr>
<tr>
<td>Dimeric isocitrate and isopropylmalate dehydrogenases</td>
<td>c.77.1.1</td>
<td>25</td>
<td>24</td>
<td>68</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>30</td>
<td>2621</td>
<td>12</td>
</tr>
<tr>
<td>Inositol monophosphatase/fructose-1,6-bisphosphatase-like</td>
<td>c.7.1.1</td>
<td>25</td>
<td>22</td>
<td>97</td>
<td>90</td>
<td>75</td>
<td>88</td>
<td>87</td>
<td>63</td>
<td>1900</td>
<td>10</td>
</tr>
<tr>
<td>ALDH-like</td>
<td>c.82.1.1</td>
<td>23</td>
<td>22</td>
<td>66</td>
<td>65</td>
<td>63</td>
<td>64</td>
<td>65</td>
<td>52</td>
<td>2704</td>
<td>11</td>
</tr>
<tr>
<td>Ferritin</td>
<td>a.25.1.1</td>
<td>23</td>
<td>20</td>
<td>88</td>
<td>85</td>
<td>82</td>
<td>81</td>
<td>85</td>
<td>54</td>
<td>1285</td>
<td>4</td>
</tr>
<tr>
<td>Class I aminoacyl-tRNA synthetases, catalytic domain</td>
<td>c.26.1.1</td>
<td>23</td>
<td>18</td>
<td>50</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>16</td>
<td>1669</td>
<td>10</td>
</tr>
<tr>
<td>ALDH-like</td>
<td>c.82.1.1</td>
<td>23</td>
<td>17</td>
<td>66</td>
<td>45</td>
<td>44</td>
<td>43</td>
<td>43</td>
<td>34</td>
<td>1284</td>
<td>11</td>
</tr>
<tr>
<td>PAPS sulfitotransferase</td>
<td>c.37.1.5</td>
<td>22</td>
<td>16</td>
<td>34</td>
<td>27</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>374</td>
<td>22</td>
</tr>
<tr>
<td>Retroviral protease (retropepsin)</td>
<td>b.50.1.1</td>
<td>20</td>
<td>19</td>
<td>220</td>
<td>219</td>
<td>215</td>
<td>216</td>
<td>216</td>
<td>199</td>
<td>1765</td>
<td>15</td>
</tr>
<tr>
<td>Thymidylate synthase/dCMP hydroxymethylase</td>
<td>d.117.1.1</td>
<td>20</td>
<td>19</td>
<td>120</td>
<td>119</td>
<td>115</td>
<td>114</td>
<td>119</td>
<td>49</td>
<td>2346</td>
<td>15</td>
</tr>
<tr>
<td>Nucleoside diphosphate kinase, NDK</td>
<td>d.58.6.1</td>
<td>19</td>
<td>17</td>
<td>60</td>
<td>55</td>
<td>53</td>
<td>54</td>
<td>52</td>
<td>31</td>
<td>799</td>
<td>25</td>
</tr>
<tr>
<td>MarR-like transcriptional regulators</td>
<td>a.4.5.28</td>
<td>18</td>
<td>15</td>
<td>18</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>2521</td>
<td>8</td>
</tr>
<tr>
<td>Ubiquitin-related</td>
<td>d.15.1.1</td>
<td>18</td>
<td>13</td>
<td>25</td>
<td>15</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>621</td>
<td>32</td>
</tr>
<tr>
<td>Glutaminase/Asparaginase</td>
<td>c.88.1.1</td>
<td>16</td>
<td>12</td>
<td>29</td>
<td>24</td>
<td>23</td>
<td>23</td>
<td>24</td>
<td>13</td>
<td>1041</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>557</td>
<td>453</td>
<td>1613</td>
<td>1422</td>
<td>1320</td>
<td>1346</td>
<td>1362</td>
<td>926</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bold numbers indicate that biological unit includes common interface in M CFGs.
When all CFGs contain an interface, when $N$ is as low as 3, 93% of entries in the PISA server contain the interface, whereas the PDB value is 78% (Fig. 6b). Generally, a larger cluster of interfaces has a lower minimum sequence identity. This also implies that those common interfaces are most likely to be biological interactions since they exist in remotely related homologues. In Fig. 6c, the lower the minimum sequence identity of proteins with the same interface in all available CFGs ($N/N$) is, the higher the percentage contained in the PDB, PQS, and PISA server is. For all values of $N$, when the minimum sequence identity is 60% or less, more than 95% of PISA server assemblies contain the interface.

**Evidence for previously contested dimer interfaces**

Comparison of crystal forms has allowed us to confirm many biological interfaces in homo-oligomeric proteins and to identify many PDB entries that are missing annotations of interfaces in existing databases. Perhaps most interesting is the possibility of identifying interfaces that have not been previously considered as part of biologically relevant structures. For 35 families with two or three crystal forms involving 145 entries, one or more common interfaces are detected in all crystal forms with a minimum identity <50%, which are not annotated or...
only rarely annotated in the PDB, PQS, or PISA server. From our benchmarking analysis, those interfaces are likely to be biologically relevant. In these cases, it is valuable to have solution experimental data that indicate that a protein is a homo-oligomer under approximately physiological conditions. We give two examples in Fig. 7.

The endosomal sorting complex I required for transport is a 350-kDa complex composed of multiple copies of Vps28 and Vps37 and a single copy of Vps23, recruited to cellular membranes during multivesicular endosome biogenesis.51 It also plays a critical role in retroviral budding.52,53 The C-terminal domain of Vps28 forms a four-helical bundle and serves as an adaptor module linked to the endosomal sorting complex I required for transport.54 The Vps28 C-terminal domain-like (a.24.28.1) family contains three CFGs and three entries (2J9W, 2J9V, and 2G3K), which share a common interface (Fig. 7a) with a minimum identity of 39% between the yeast and *Xenopus laevis* structures. The full-length Vps28 can form homodimers in solution,54 whereas the C-terminal domain alone is monomeric. However, the common interface in three CFGs with a sequence identity of 39% indicates that the C-terminal domain may also form a homodimer, probably stabilized in the presence of

![Fig. 6. Percentages of entries in the PDB, PQS, and PISA server that contain common interfaces. (a) Percentage of entries in the PDB, PQS, and PISA server versus M (M = 2, 3, 4, 5, 6, 7, 8, 9, 10, >10) and M/N ≥ 0.5. (b) Percentage of entries in the PDB, PQS, and PISA server versus N (N = 2, 3, 4, 5, 6, 7, 8, 9, 10, >10). (c) Percentage of entries in the PDB, PQS, and PISA server versus minimum sequence identity ranges for N/N interfaces.](image)

![Fig. 7. Examples of interfaces not well annotated in public databases. (a) 2J9W interface (entry A: X,Y,Z; entry B: X,Y,Z). (b) 1EAQ interface (entry A: X,Y,Z; entry B: 1/2 – X,1/2+Y,1–Z).](image)
the homodimeric N-terminal domain. The PDB and PQS do not have the interface in the biological units for any of these three entries, whereas the PISA server does for entries 2J9V (yeast Vps28) and 2J9W (Xenopus Vps28), but not 2G3K (yeast Vps28). For 2G3K, the PISA server has a different dimer, with the monomers in a perpendicular orientation to each other.

The Runt homology domain is an evolutionarily conserved DNA-binding domain and is essential for hematopoiesis. In our results, in the family Runt domain (b.2.5.6), there exists an interface (Fig. 7b) formed by two Runt domains with three CFGs containing five entries. The interface is composed of a two-stranded parallel β sheet consisting of strands β3 and β12 as well as parts of loops L3 and L12 of the Runt domain. The interface is similar to the interface reported in STAT proteins, which includes strands βa’ and βg’ and the succeeding loops connecting βa’ with βb and βg’ with helix α5. These loops participate in DNA binding. Moreover, one of five buried water molecules, W1, is part of a major groove-binding motif that is conserved in Runt and STAT proteins. Only PQS contains this dimer in the biological unit of one of these PDB entries (1EAQ). The PDB and PISA server do not have the interface in any of these four entries.

**Discussion**

The structures of oligomeric assemblies of proteins are important for understanding their functions and regulation and the phenotypes of mutations. At the present time, there exist few repositories of experimental data on the oligomeric state of proteins in solution, such as gel filtration, analytical ultracentrifugation, and other experiments, to determine the size of protein assemblies under even approximately physiological conditions. Even these experiments may not cover the range of physiological conditions under which a protein may change its oligomerization state and therefore may fail to identify some physiologically relevant states. Two databases that are available are Doodle and BRENDA. Doodle presents data on homotypic interactions derived by experiments using lambda repressor fusions. Proteins with self-association (homodimers and larger) will provide resistance to lambda phage infection. The experiments have been performed on the *Escherichia coli* and yeast genomes. Another useful database is BRENDA, which collects information on enzymes based on enzyme classification assignments. Through BRENDA, it is often possible to find experimental data on the number of subunits in some enzyme assemblies. However, even here, the experiment may be performed on the enzyme from one organism while the structural information may be available from a different organism. Although some proteins are “well known” to be dimers or tetramers, the experimental data on which those are ultimately based are often obscure. This is a common problem in other areas, such as gene ontology, in which annotations are copied from gene to gene by reason of homology, but the original experiment may be hard to track down.

Even when the size of an oligomeric assembly is known, it is often unclear what the correct structure for that assembly is purely from X-ray crystallographic structures. The authors of crystal structures provide annotations for hypothetical biological units, but the reasoning that leads to these annotations is completely undocumented in the PDB itself, and often missing in the literature as well. There are several lines of evidence commonly presented in articles on crystal structures for what represents the biological structure. Often the largest interface is assumed to be biological, especially if it is C2-symmetric. This is often, but certainly not always, correct. In addition to surface area and symmetry, higher residue conservation in one interface compared with another may indicate those more likely to be biological. Mutation data that indicate residues likely to be in a biological interface may also be available.

In this work, we have extended a method used by many crystallographers to determine biological interactions among identical subunits—the comparison of multiple crystal forms of a protein or members of a protein family. Some proteins have been crystallized in a surprisingly large number of crystal forms, and some dimer interfaces have been observed in dozens of different such forms. It is very likely that an interface observed in a large number of crystal forms is biological. This is especially the case if the interface is observed for different family members, because it is less likely that crystal contacts would be very similar for distantly related proteins than for identical sequences. Indeed, we found for PDB entries in a benchmark constructed from the data of Bahadur et al. and Ponsinogl et al. that monomeric proteins rarely have common interfaces in more than one-third of their crystal forms and usually much less, whereas multimeric proteins have common interfaces in a majority and frequently all their available crystal forms. NMR structures confirm these results.

As more structures are determined in large protein families, the comparison of crystal forms may play an important role in suggesting which interfaces are important biological interactions. In subsequent work, we will present software and a database for comparison of interfaces in new structures with those in existing structures of homologous proteins, so that this form of structural reasoning will be readily available to crystallographers. Such analysis is difficult to perform manually and visually for anything more than a few crystal forms. We are also extending this work to heterodimeric complexes and multidomain proteins.

Finally, we believe our data set of interfaces observed in a large number of crystal forms will provide computational researchers with sets of interfaces that can be used as training and testing data for predicting biological interfaces in new structures and for docking calculations. Currently, such methods are based on interfaces assumed to be true because they are in
the PDB and PQS. Our work has demonstrated that these databases certainly contain incorrect interfaces and at the same time are missing many interfaces that are observed in large numbers of crystal forms. In the absence of a large data set based on cross-linking and/or NMR data, interfaces observed in a large number of crystal forms will remain those most likely to be correct and therefore the most reliable set for training and testing of prediction methods.

Materials and Methods

Data sources

The data used in this study come from five sources: protein structure files from the PDB in XML format (PDBML)\textsuperscript{19,60}; biological unit coordinate files from PQS\textsuperscript{20} in the legacy PDB format; PISA server multimers in XML format\textsuperscript{21}; domain classification files from SCOP 1.73\textsuperscript{22}; and CE/PSI-BLAST hit files from a nonredundant (100%) PDB database in our laboratory.\textsuperscript{84,65} We built a crystal form and PDB biological unit from the asymmetric unit information given in the PDB XML files. We used the most recently remediated XML files released in beta form in April 2007 and in final form on August 1, 2007\textsuperscript{1}.

Classifying PDB homomultimers

Families

We used SCOP 1.73 to define families of PDB entries, using only PDB entries with a single protein sequence and a single SCOP family designation (e.g., a.1.1.1). Only those SCOP families with class designation from a to g were used. We added to these families all single-sequence entries deposited into the PDB since the last data set included in SCOP 1.73, if the newer entries had a PSI-BLAST E value better than 0.001 with the SCOP-defined entry and either a sequence identity >50% and an alignment length >80% of both proteins or a CE structure alignment Z score ≥4.5. The PSI-BLAST and CE data were taken from our PISCES database.

Crystal forms in each family

We initially divided each family into crystal forms, which are subsets of entries in the family with the same space group, the same number of monomers in the asymmetric unit, and similar crystal cell dimensions and angles (±1%). These initial crystal forms were combined into CFGs based on a comparison of interfaces in each crystal form (see Results).

Representative entries for each crystal form

We used the entry in each crystal form with the highest X-ray resolution as the representative entry of the crystal form to compare interfaces between two crystal forms.

Building crystal structures

A crystal structure was built from the asymmetric unit and space group defined in the XML PDB files. If non-crystallographic symmetry operators are given in the XML file and the coordinates for these symmetry copies are not given, the operators were used to build the asymmetric unit first. The symmetry operators for forming each PDB biological unit in Cartesian (Angstrom) form were provided by the PDB (Z. Feng, personal communication).

Building crystals using space group symmetry operators

The scale matrix is read from the PDB XML file for each entry:

\[
S = \begin{pmatrix} s_{11} & s_{12} & s_{13} & a_1 \\ s_{21} & s_{22} & s_{23} & a_2 \\ s_{31} & s_{32} & s_{33} & a_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\]

The fractional coordinates are therefore:

\[
X'_f = SX
\]

where \(X = (x, y, z)^T\), which are the atomic Cartesian coordinates of an atom.

Each space group has its own symmetry operators, which are applied to the asymmetric unit to generate enough symmetry copies to build a unit cell. The number of symmetry operators ranges from 2 to 96. A symmetry operator is defined as

\[
P_m = \begin{pmatrix} p_{11} & p_{12} & p_{13} & t_1 \\ p_{21} & p_{22} & p_{23} & t_2 \\ p_{31} & p_{32} & p_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\]

Here, \((t_1, t_2, t_3)^T\) is the translation vector. A fractional coordinate is transformed by \(P_m\) to \(X_m\):

\[
X_m = P_m X_f = P_m SX
\]

Applying the standard symmetry operators results in some copies of the asymmetric unit either in part or entirely in a neighboring unit cell or even two or more cells away. We constructed a unit cell first by scaling the coordinates of the asymmetric unit and calculating its geometric center. For instance, the geometric center of the original asymmetric unit is \((0.01, 0.76, 2.41)\) in the scaled coordinate system. This is therefore the unit cell with coordinates between \((0, 0, 2)\) and \((1, 1, 3)\). We then applied the symmetry operators to this structure, and if the geometric centers of other asymmetric units place them outside the \((0, 0, 2)\) unit cell, we translated these symmetry copies into this unit cell by adding \(±1, ±1, ±1\), etc., as necessary.

To identify contacts among monomers in the crystal, we built the complete unit cell and then copies of this unit cell in both the plus and minus directions in one, two, and three dimensions. This resulted in a \(3 \times 3\) collection of unit cells. These translations are formed from the scaled unit cell coordinates by adding \(1\) or \(-1\) to the coordinates:

\[
X_{\text{unit}} = X_{\text{unit}} + T_r
\]

where \(T_r = (i, j, k)^T\) and \((i, j, k)^T\) is the \(r\)th vector in the following series: \((-1 -1 -1)^T, (-1 -1 0)^T, (-1 -1 1)^T, \ldots\), \((1 +1 +1)^T\).

The final stage is to convert fractional coordinates to Cartesian coordinates by multiplying by the inverse of the scale matrix:

\[
X_m' = S^{-1} X_m = S^{-1} (P_m S X + T_{\text{unit}})
\]

Identifying interfaces in a crystal structure

The number of atomic distances that must be calculated to identify interfaces within crystals is reduced by using a

\[\text{http://www.pdb.org}\]
tree of bounding boxes. The method is based on \(k\)-dimensional discrete oriented polytopes.\[^6\] These are multifaceted bounding boxes with faces perpendicular to one of \(k\) fixed axes. We used \(k=3\), corresponding to the three Cartesian axes. A \(k\)-dimensional discrete oriented polytope is a hierarchical tree of bounding boxes with the largest box around an entire protein and leaf nodes around each residue of the protein. The procedure to build a tree using a top–down method is as follows:

1. Calculate the geometric center of each residue.
2. Calculate a bounding box for all residues of the entire protein as the tree root.
3. Select the axis with largest variance of the current bounding box.
4. Split the tree into two branches by the mean of coordinates along the selected axis.
5. Calculate the bounding box for each branch.
6. Check if either branch is a single residue. If so, it is a leaf node, with a bounding box containing the atoms of the residue. Otherwise, go to step 3.

We built a tree for each chain in the unit cell containing the original asymmetric unit. The same tree was used for all other unit cells to save time and computer memory. However, because of the translations and rotations, the bounding boxes for these unit cells had to be rebuilt, starting from the leaves of the tree and working our way up to each tree root (i.e., each protein chain). All atomic contacts were calculated whenever there were overlaps between two leaf nodes.

### Comparing interfaces

To compare interfaces between two crystal forms, we compared only unique interfaces from one representative entry of each crystal form. An interface is defined between two chains if the chains have at least 10 amino acid contacts and at least one atomic contact. Here, an amino acid contact is defined as two \(C^\alpha\) atoms \((C^\alpha\text{ for glycine})\) with a distance less than 12 Å. An atomic contact exists between residues with at least one atomic distance less than 5 Å. The interfaces of an entry are considered to be the same (or nonunique) when they consist of the same pair of chains from the asymmetric unit and have a \(Q\) score greater than 0.95 (defined below). For instance, an interface is a symmetry copy of another interface if both protein pairs are made of asymmetric unit chains A and B, but a similar interface made from two copies of chain A would be different.

Interface similarity between two homologous entries was measured with the score function \(Q\) described in our earlier article.\[^{34}\] The score \(Q\) reflects the similarity of contacts between two protein dimers. Because the proteins in the two dimers may be homologous, rather than identical, we sought to derive a score function that was not sensitive to the specific residues in the interface. That is, two homologous dimers may be highly similar in orientation but the existence of atomic contacts between corresponding residues (as derived from sequence or structure alignment) may be different because of the substitution of longer for shorter side chains and vice versa. We therefore required a method that is not sensitive to side chain identities in the interface but rather whether the two protein backbones in one dimer have roughly the same orientation as the backbones in the other dimer. \(Q\) is therefore defined based on comparing the \(C^\alpha\text{--}C^\alpha\) distances of corresponding amino acids in the two interfaces (we use \(C^\alpha\) for glycine in what follows). It is a weighted sum of differences in distances of corresponding backbone atoms in the two interfaces. If we define the two distances as \(e_i\) and \(f_i\) and the weight \(w_i\) is some monotonically decreasing function of \(d_i = \min(e_i, f_i)\), then \(Q\) is defined:

\[
Q = \frac{\sum w_i \exp(-0.5|e_i - f_i|)}{\sum w_i}
\]

The similarity function uses a distance-weighted score with a weight of 0 for a \(C^\alpha\text{--}C^\alpha\) distance greater than or equal to 12 Å (\(C^\alpha\) for Gly). This value was selected by calculating the \(C^\alpha\text{--}C^\alpha\) distances for a large set of protein dimers from the PDB and measuring the probability of an atomic contact (\(\leq 5\) Å) if the \(C^\alpha\text{--}C^\alpha\) distance was less than \(D\). As shown in Fig. 8a, this probability, \(p(\text{atom contact} | \text{dist} (C^\alpha) - \text{dist} (C^\alpha) \leq D)\), goes to nearly zero when \(D\) is 12.0 Å. This probability function is fit well to a function \(f\) that consists of half of a Gaussian distribution \(N(5, 2.28)\) when \(D \geq 5\) Å and 1.0 if \(D < 5\) Å, as shown in Fig. 8b. Therefore, the weight function is defined:

\[
y = \begin{cases} 
1 & 0 < x \leq 5 \\
2.08 \exp \left( \frac{-(x - 5)^2}{5.139} \right) & x > 5 
\end{cases}
\]

In this work, we have used this weight function. In our earlier work, we used a quadratic function similar to the switching function used in CHARMM for turning off nonbonded interactions above some distance.\[^67\] The results with the two functions are very similar, but the new function can be derived more directly from protein structures.

To determine the minimum \(Q\) score for two possible similar interfaces, we selected a list of homodimers with \(<20\%\text{ sequence identity with each other and calculated the minimum }C^\alpha\text{ distances for each }C^\alpha\text{ in the interface. The distribution of }C^\alpha\text{ distances fits into a Gaussian distribution (Fig. 8c).}

\[
y = 0.036e^{\frac{(x - 17.31)^2}{272.38}}
\]

From this Gaussian function, we generated 1000 pairs of distances with size 300 and calculated \(Q\) scores. The \(Q\) scores are also fit into a Gaussian distribution (Fig. 8d):

\[
y = 0.114e^{\frac{(x - 0.026)^2}{0.0339}}
\]

When the \(Q\) score was \(>0.08\), the probability for the \(Q\) score from two arbitrary dimers is close to 0. In our study, we set the minimum \(Q\) score to be 0.1.

### Surface area

The surface area of each unique interface in each crystal form was calculated with the program NACCESS.\[^{35}\] The interface area is the sum of the surface areas of the two individual proteins minus the surface area of the protein complex divided by 2:

\[
\text{Interface Surface Area} = (\text{SASA}_A + \text{SASA}_B - \text{SASA}_{AB}) / 2
\]

### Clustering family interfaces

We clustered interfaces with a surface area greater than 200 Å\(^2\) that occur in at least two crystal form representatives using a hierarchical clustering algorithm. Interfaces
Fig. 8. (a) Relationship between residue distance and contact probability. (b) Gaussian fit of (a). The data on the left part of (b) were mirrored from the right part in order to model the data. (c) Distribution of $C_{\beta}$ distances, which are calculated from a list of homodimers with <20% sequence identity with each other. (d) Distribution of $Q$ scores computed from 1000 pairs of distances with size 300 from the Gaussian function in (c). Gaussian fits were modeled by Original Plot 7.0.
in a family are sorted by the number of similar interfaces. The clustering started from the interface with the largest number of similar interfaces. An interface was added into a cluster C only if it has a Q > 0.10 with at least half of the interfaces already in cluster C.

Benchmark data set

The benchmark data set was compiled from the Pons-tingl et al.33 and Babadur et al.36,37 data sets in our study to verify if common interfaces in multiple CFGs correlate with biological interactions. Only those entries with single protein sequence and one single SCOP domain per chain were selected. The reference data set contains 258 entries in which there are 132 monomers and 126 homomultimers.

Implementation

The program is written in C#.Net. Data are stored in a Firebird relational database®. The database structure was designed to be modular, to avoid unnecessary redundancy and to allow fast queries. Our database can be divided into independent modules: SCOP, PDB, PQS, PISA, family definition, structure alignment, crystal interfaces definition including symmetry operators, asymmetric chains, atomic contacts, residue contacts and distances, interface comparison, and crystal interface and PDB/PQS/PISA biological unit interface comparison. All interfaces that occur in at least two crystal forms are output as PDB formatted files.

Availability

Structures of protein dimers observed in multiple crystal forms are available from us online® (250 MB). Phylogenetic trees of SCOP families with common interfaces are also available online® (2.5 MB). Other supplementary materials are available from the journal Web site.

Acknowledgements

This work was supported by the National Institutes of Health through grant R01 GM73784 awarded to R.L.D. We thank Dr. Longin Jan Latecki for useful comments. We also thank Zukang Feng and Eugene Krissinel for providing information required to build PDB and PISA server biological units, respectively. Rajib Mitra and Brian Weitzner assisted by visually examining many PDB structures to verify the ability of the parameter Q to identify similarities and differences in interfaces and crystal forms.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.06.002

References


http://firebird.sourceforge.net/
http://dunbrack.fccc.edu/JMB08/InterfaceFiles.tar
http://dunbrack.fccc.edu/JMB08/SupplementaryFigures.tar


