Published in final edited form as:
J Mol Biol. 2008 August 29; 381(2): 487–507. doi:10.1016/j.jmb.2008.06.002.

Statistical analysis of interface similarity in crystals of homologous proteins

Qifang Xu[1,2], **Adrian A. Canutescu**[1], **Guoli Wang**[1], **Maxim Shapovalov**[1], **Zoran Obradovic**[2], and **Roland L. Dunbrack Jr.**[1*]

1*Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia PA 19111, USA*

2*Center for Information Science and Technology, Temple University, 1805 N. Broad Street, Philadelphia, PA 19122, USA*

Abstract

Many proteins function as homooligomers and are regulated via their oligomeric state. For some proteins, the stoichiometry of homooligomeric states under various conditions has been studied using gel filtration or analytical ultracentrifugation experiments. The interfaces involved in these assemblies may be identified using crosslinking and mass spectrometry, solution-state NMR, and other experiments. But 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 PDB entries in a protein family reveals several important features. First, similarity of space group, asymmetric unit size, and cell dimensions and angles (within 1%) does not guarantee that two crystals are actually the same crystal form, that is containing similar relative orientations and interactions within the crystal. Conversely, two crystals in different space groups may be quite similar in terms of all of the interfaces within each crystal. Second, NMR structures and an existing benchmark of PDB crystallographic entries consisting of 126 dimers and larger structures and 132 monomers was used to determine whether the existence or lack of existence 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, while 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 PISA database available from the EBI is more consistent in identifying interfaces observed in many crystal forms than is the PDB or EBI'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.

Introduction

Many proteins are oligomeric due to the association of identical subunits under physiological conditions. Homooligomerization may be part of allosteric regulation[1], or contribute to conformational and thermal stabilities[2] and to higher binding affinity with other molecules. Homodimeric proteins have been found to form interactions with a larger number of other

*Correspondence: Roland.Dunbrack@fccc.edu.

proteins than monomeric proteins[3]. Multimerization is particularly common in enzymes, transcription factors, and signal transduction[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[5; 6]. Some proteins oligomerize by domain swapping, in which a segment of monomeric protein is replaced by an identical segment from the other subunit and vice versa[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 to dissimilar interfaces. Heterodimers of proteins in the same superfamily may then evolve from such homodimers.[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 α1(I) chain, producing abnormal disulfide-bonded dimeric α1(I) chains[10]. Myofibrillar myopathy (MFM) is a human disease of muscle weakening, 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[11]. Cu,Zn superoxide dismutase (CuZnSOD) is an efficient enzyme that catalyzes the conversion of superoxide to oxygen and hydrogen peroxide[12]. Familial amyotrophic lateral sclerosis or Lou Gehrig's disease is associated with mutations in CuZnSOD[13; 14]. Some mutations destabilize the SOD dimer, causing abnormal aggregation that may be lethal to cells.

Experimental means for determining the size of an oligomer include analytical ultracentrifugation[15] and gel filtration[16]. These methods separate proteins and protein complexes based on their size or mass, from which 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[17]. Fluorescence resonance energy transfer (FRET) 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[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, information on oligomeric association size and in particular structure comes from X-ray crystallography. For many proteins the size and actual structure of multimers is controversial or unknown, and is based *only* on what is observed in crystal structures, sometimes even a single crystal structure. Both the Protein Data Bank (PDB)[19] and the European Bioinformatics Institute (EBI)[20; 21] provide information on "biological units" or assemblies which 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, while those of the recently developed PISA server[21] from the EBI are based on the analysis of interfaces and predicted stability of complexes observed in single crystal structures. PQS (Protein Quaternary Server)[20] 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, while PISA contains multiple oligomeric structures of different sizes for many PDB entries based on chemical thermodynamics calculations on complex stability. The recently developed PIQSI database provides manually annotated sizes of biological units from the literature for PDB entries[22].

Many databases and analyses have used PDB and PQS biological units to examine the interfaces between protein domains. For instance, PIBASE[23] provides a list of structures for a query of two SCOP superfamily or family designations[24], and provides access to coordinates for each pairwise interaction. Interactions in 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 non-physiological interactions, and hypothetical biological units as proposed by the authors of PQS, (e.g., 1ylv.mmol). The emphasis is on characterizing pairwise interfaces in terms of surface area and polar/nonpolar content. PSIMAP/PSIBASE[25] 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[26], SCOPPI[27], and iPFAM[28] also use SCOP, PFAM, PDB and PQS to define atomic interactions among protein domains. Databases of this sort are used for statistical analysis of residue contacts across interfaces to develop methods for predicting or scoring interfaces[5; 29; 30; 31; 32; 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[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 unwarranted assumption that the asymmetric and biological units are the same. We also found that 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 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 multimer 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[34].

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[35], and compared the ability of buried surface area and pair interaction scores to predict biological contacts in crystals. This dataset has subsequently used by others as a benchmark for methods that attempt to determine biological assemblies from single crystals[21; 22]. Bahadur et al.[36; 37] assembled a set of interfaces consisting of 70 heterodimeric structures, 122 homodimeric structures, and 188 crystal packing interfaces with surface area greater than 800 $Å^2$, and examined the physical properties of the different interface classes. Shoemaker et al.[38] 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 paper, we examine thoroughly the interfaces in crystals of single homologous proteins. We attempt to answer several questions. First, when are two crystals of the same or similar proteins really the same crystal form and when are they not? We find surprisingly 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 of the interfaces within each crystal. This occurs when one contains a subset of symmetry operators of the other and a larger

asymmetric unit, and also when one is a small distortion of the other such that the space group is different. This analysis helps to sort PDB entries within a family into truly different crystal forms.

Second, we examine 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 compare all interfaces in the available crystal forms in each family and determined those shared by two or more crystal forms. We determine the number of crystal forms with the interface, *M*, compared to the total number of different crystal forms in the same family, *N*. We then evaluate the usefulness of these numbers with prior benchmarks on oligomeric interactions as well as 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 find 36 families in which all *N* out of *N* crystal forms contain a particular interface, where $N \geq 10$. These interfaces are very likely to be physiological. We also find that monomers in a benchmark set comprising both the Ponstingl and Bahadur sets tend to have $M << N$.

Third, we examine 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 2 or 3 such forms and are not shared by homologous proteins. That is, they are only formed under non-physiological crystallization conditions including high protein concentration, peculiar pH, and the presence of non-physiological ligands. This has previously been observed for T4 lysozyme, which has been studied in many crystal forms[39]. When an interface is shared in two different crystal forms by divergent proteins, then the interface is very likely to be biologically important. We also find that in large families, some interfaces 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. 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. PISA 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. PISA is therefore the best source of biological assembly information when only one or two crystal forms is or are currently available.

This study is closest to the work of Shoemaker et al.[38] but with some important differences. First, we examine the interfaces across PDB entries of homologous proteins to determine whether they are or are not the same crystal form, despite similarities and differences in space group, asymmetric unit size, and unit cell dimensions and angles. Shoemaker et al. separate crystal forms only by space group and/or differences in cell dimensions of greater than 2%. We find that this is inadequate to classify crystals as similar or different. Second, we evaluate the usefulness of the number of different crystal forms and the evolutionary relationships of shared interfaces, neither of which is considered by Shoemaker et al. Finally, we provide in supplemental 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 homooligomeric structures, and so only PDB entries with a single polypeptide sequence and no nucleic acids 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 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 symmetry operators of the other. When looking at different crystals of the same protein or related proteins, we have observed many PDB entry pairs in different space groups that when viewed with molecular graphics, are observed to be essentially the same crystal form. As an example, we show in Figure 1 (top) the crystal lattice of PDB entries 1IJV[40] and 2NLQ[41], the protein defensin, with different space groups, P 2_1 2_1 2_1 and C 1 2 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), same space group, the same asymmetric unit size, and unit cell dimensions and angles within 1% of each other. There were a total of 2,991 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 Methods) of the function Q that we have developed previously to compare interfaces in different pairs[34]. Q expresses the fraction of interacting residues that are similar between two interfaces, and is similar to scores used to judge docking success in the CAPRI experiments[42]. A value of Q=1 means that all contacts in one interface are shared in the other interface and vice versa. A value of 0 means that no contacts are shared. Random simulations indicate that a value of $Q \geq 0.1$ indicates a statistically significant similarity of two interfaces (see Methods).

We compared crystals using different minimum values of Q and minimum values of 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 surface area $\geq ASA_{min}$ and compared each interface with *all* of the interfaces in the second member of the pair (entry B), regardless of surface area in entry B. The reason for this is that an interface with area just over ASA_{min} in A might correspond to an interface in B with area just below ASA_{min}., which would be missed if ASA_{min} were applied to B. We calculated f_{AB} as the fraction of interfaces in A with surface area $\geq ASA_{min}$ that have interfaces in B with $Q \geq Q_{min}$. If f_{AB} =1, then all interfaces in A have corresponding interfaces in B given the ASA_{min} and Q_{min} cutoffs. This process was then performed in reverse (interfaces of B with $ASA \geq ASA_{min}$ against all interfaces in A) to calculate f_{BA} given the same cutoffs.

In Table 1, we show the number of same-sequence entry pairs for which both f_{AB} =1 and f_{BA} =1 for various values of ASA_{min} and Q_{min}. First, for Q_{min} = 0 and larger minimum surface areas the number drops below 2,991, because a small number of crystals have no interfaces larger than those values of ASA_{min}, and therefore f_{AB} and f_{BA} are undefined. For values of $Q_{min} > 0$, the number of pairs with f_{AB} =1 and f_{BA} =1 drops off at lower values of ASA_{min}. We examined visually the 7 entry pairs that do not have f_{AB} =1 and f_{BA} =1 with Q_{min} = 0.1 and ASA_{min}=100 $Å^2$. We found 4 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. While 1LMN and 1LMC are visually the same crystal as are 1BB6 and 1BB7, 1LMN and 1LMC are different from 1BB6 and 1BB7. The other 3 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 the one hand, and 1BB6 and 1BB7 on the other). This resulted in 919 entry pairs for comparison using various cutoffs of ASA_{min} and Q_{min} as shown in Table 2. This time, for $ASA_{min} \geq 250$, some entry pairs have no interfaces above ASA_{min} and so f_{AB} and f_{BA} are undefined. We examined visually the 31 entry pairs for which either f_{AB} or $f_{BA} \leq 1$ when $ASA_{min}=100$ and $Q_{min}=0.1$. 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 out of 31, PDB entries 1DPO (anionic trypsin) and 1ZSL (Factor XIa) (39% sequence identity), are different crystals. These entries are the same space group (I 2 3), asymmetric unit size (monomer), and 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 value of $Q_{min}=0.1$, the highest detection rate of similar crystals, i.e., with both $f_{AB}=1$ and $f_{BA}=1$, occurs for $ASA_{min}=200$ Å^2 (in fact at all values of $Q_{min} > 0$ this is true). If we define false negatives as crystals which are actually the same crystals but which fail these criteria, with the five different crystal entry pairs described above (true negatives), these cutoffs contain only 14 false negatives (2991 + 919 − 2986 − 905 − 5 = 14) out of 3,910, 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 $f_{AB}=1$ and $f_{BA}=1$, we found no false positives, but it is difficult to calculate a true false positive rate. Given the same space group, 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 different unit cell dimensions. Using cutoffs of when $ASA_{min}=200$ and $Q_{min}=0.1$, we found 6,931 pairs (of 276,643) with $f_{AB}=1$ and $f_{BA}=1$. Most of these pairs (90%) had high sequence identity (≥90%). A total of 1,375 pairs were 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 of the interfaces in one crystal with ASA≥200 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 back-to-front in one crystal while back-to-back in the other. An example of this is shown in Figure 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 of these planes is 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 together that contain similar non-biological interfaces, those that would produce infinite lattices, we decided to group together entries that have $f_{AB}=1$ and $f_{BA} \geq 0.65$ or $f_{AB} \geq 0.65$ and $f_{BA}=1$. This occurs for a further 2,883 entry pairs for the

representatives described above. The value of 0.65 was chosen after inspecting many entry pairs. If we group entries that all 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 8,816 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 8,816 groups, *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 crystal form groups is shown in Figure 2.

Common interfaces in multiple crystal form groups

Some statistical information on these CFGs is shown in Table 3. There are 1,125 SCOP-defined families with at least two CFGs. Most CFGs contain highly similar sequences, with only 5% of CFGs with minimum sequence identity < 90% (Figure 3, top left) while between any two CFGs in a family, 95% of CFG pairs have minimum pairwise sequence identity < 90% (Figure 3, top right). The number of crystal forms groups in a family ranges from 1 to 173 (SCOP family *protein kinases, catalytic subunit*, d.144.1.7), and 51% of families have 2 or 3 CFGs (Figure 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 best resolution were compared to 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 value of Q_{min}=0.1. Table 1 also gives the overview of the total common interfaces. A total of 868 families have at least one common interface in two or more crystal form groups, involving 15,264 entries and 3,771 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 (Figure 3, bottom right). There are 176 families with at least one common interface in all crystal form groups with $N \geq 4$, which comprise 1,372 CFGs and 3,139 entries. There are 248 families containing 2,136 CFGs and 5,200 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. The percentages of common interfaces available in PDB, PQS and PISA biological units are also shown in Table 3. For those interfaces in a large number of crystal forms, PISA identifies 97% of these interfaces as part of a biological assembly when $M=N$ 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] have 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 existing databases of the PDB, PQS, and PISA. These results are shown in Table 4. While 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 PDB and PISA and more than half the time by PQS. This is mostly due to the monomers in the Bahadur set, which were chosen to contain crystallographic interfaces larger than 800 $Å^2$.

We first examined the monomers in the benchmark, and determined whether any of the interfaces in each crystal appeared in other crystal form groups 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 appears in according to a similarity criterion of $Q \geq 0.1$. A bar chart of M vs. N for the 132 monomers is shown in Figure 4a. Only 4 of the entries are in families with only one CFG. Of the remaining 128 entries with 2 or more CFGs, 55 of them 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 two entries in the benchmark with M=14 are both T4 lysozyme, and the associated interface is one previously noted as occurring in many crystal forms by Matthews et al.[39] A structure of alpha bungarotoxin, 2ABX[45], has a common interface with M = 9. This interface also exists in the NMR structure of neuronal bungarotoxin (45% identity), PDB entry 2NBT, although this structure was solved partly by homology to crystal structures since there were few interprotein NOEs[46]. PDB and PQS have the same dimer, but the dimer from PISA contains a different interface.

We examined the 126 dimers and higher oligomers in the benchmark, and determined whether these crystals contain interfaces observed in multiple crystal form groups. A plot of M vs N for the largest M for each entry is shown in Figure 4b. The distribution is quite different from that of the monomers. A total of 55 (44%) of the benchmark oligomers contain at least one interface that is present in all N out of N CFGs in each family. Seventy-seven (77) 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, but only the size of the assembly in solution experiments, we cannot be sure that all of these interfaces are biological. Nevertheless, the monomer and oligomer benchmark data shown in Figure 4(a, 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/Bahadur set does not contain coordinates for interfaces involved in the dimers and larger oligomers. So we examined the values of M and N for interfaces found in families for which there is at least one NMR structure. There are 598 such families. For these families, we looked at the crystal form groups and determined M and N for the interfaces in 2 or more CFGs, and then determined whether these interfaces were in any available NMR structures for the family. The results are shown in Figure 4c and 4d and divided into those interfaces not found in NMR structures (Figure 4c) and those that are found in at least one NMR structure (Figure 4d). The figures are quite similar to the Ponstingl/Bahadur monomers (Figure 4a) and dimer/oligomers (Figure 4b), respectively. When an interface is not in any NMR structure, M tends to be quite low and $M << N$; i.e., there are few points near the diagonal at larger N. We examined the eight cases in Figure 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 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 materials.

In order to derive a probability that an interface is biologically relevant from M and N, P (*Biological interface* | M, N), we combined the benchmark data set described above and the PIQSI database[22]. PIQSI is a manually curated set of biological unit sizes for 15,000 entries in the PDB. These were derived from literature references as well as comparison across PDB entries within families. PIQSI 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 4,171 monomers and 3,090 dimers/oligomers. Plots similar to Figure 4 for monomers and dimers/oligomers in PIQSI look very similar to Figures 4a and 4b respectively, although with somewhat more noise (data not shown). Further, 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% (2,185 monomers and 254 dimers/oligomers) and a data set with identity < 90% (1,986 monomers and 2,836 dimers/oligomers). If an interface only exists in

one CFG, the minimum identity is set to be 100%. The probability P(*Biological interface* | M, N) is calculated by

$$P=\frac{\#\text{oligomers}(M,N)}{\#\text{monomers}(M,N)+\#\text{oligomers}(M,N)}$$

for each M,N combination, where "#oligomers" includes both dimers and larger oligomeric structures.

Table 6 shows the probabilities for each M and N from the data set with identity ≥ 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 Table 6 and Table 7 are provided in the supplementary materials. When proteins share high sequence identity (>90%), there is some tendency for similar interfaces to be used in different crystal forms even in cases where the proteins are likely to be monomers under approximately physiological conditions. However, when there is some sequence divergence (≤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 crystal form groups

We examined families for which *all* available crystal forms groups contain one or more common interfaces (Figure 3, bottom right). We found 36 families (594 crystal forms, 1,585 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 Supplemental Material. The interfaces described in Table 6 are almost certainly biological interactions, since it is highly unlikely that a non-biological 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, while PQS biological units contain the interfaces in 94% of the entries. For PISA, we take the most stable assembly defined for each PDB entry; PISA 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 Å^2 to 4800 Å^2. Some of the families contain more than one shared interface, implying the existence of common tetramers and larger structures. For instance, the *L-aspartase/fumarase* family (SCOP a.127.1.1) has 3 common interfaces.

Some SCOP families are broader than others, containing relatively distant homologues and clear paralogues. At larger evolutionary distances, we expect some biologically relevant interfaces will not be conserved or will have evolved in one branch of a family but not others[47]. It is always possible that some crystals form under conditions under which dimers and larger structures do not form, or under which only monomers enter the crystal[18]. We therefore examined families for which M out 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, 1,437 entries) with common interfaces in M out of N crystal forms, for which $N \geq 15$ and $M/N \geq 0.7$. PDB, PQS and PISA contain those common interfaces in 93%, 95% and 96% of the entries respectively. A table with $M \geq 4$ and $M/N \geq 0.7$ is given as Supplementary Material.

For many of these families, phylogenetic trees show that the interface is conserved within an entire branch of the family. We give *ribonucleotide reductase-like* family (SCOP a.25.1.2) as an example. Members of this family have been demonstrated to be homodimers[48; 49; 50] in gel electrophoresis experiments. A phylogenetic tree of the PDB entries in this family is shown in Figure 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 average accessible surface area of 1985.3 Å^2. In the main part of Figure 5,

those entries that contain this interface are individually boxed. Entries are colored by their crystal form group. The PDB has the common interface in all its biological units. PQS has 3 monomers (2O1Z, 1SYY, 2ANI), and PISA has one monomer (1SMQ) and one dimer (2P1I) that is not the common dimer in the 11 crystal forms. The other common interface (Figure 5, inset) is observed in the stearoyl acyl-carrier desaturases in 4 crystal forms (7 entries) with surface area 2722.2 $Å^2$. These 7 entries are remotely related to the entries in cluster 1 with identity less than 10%. PDB, PQS and PISA all give the same biological dimers with the common interface. The entry 1OTK, phenylacetic acid degradation protein (PAAC) is distantly related to all other entries in this family with sequence identity less than 10%. No common interface was detected between 1OTK and other structures in this family.

We provide images of the phylogenetic trees for the families listed in Table 9 similar to that in Figure 5, with those entries containing a common interface marked in color, in this file: http://dunbrack.fccc.edu/JMB08/SupplementaryFigures.tar.

Statistical analyses of common interfaces in PDB, PQS and PISA

We further analyzed the tendency for PDB biological units, PQS quaternary structures and PISA's most stable assemblies to contain interfaces found in multiple crystal form groups (Figure 6). The larger the value of *M* is the more likely that the PDB, PQS, and PISA biological units will contain an interface when $M/N \geq 0.5$ (Figure 6a). When all crystal form groups contain an interface, when *N* is as low as 3, 93% of entries in PISA contain the interface, while the PDB value is 78% (Figure 6b). Generally, a larger cluster of interfaces has a lower minimum sequence identity. This also implies those common interfaces are most likely to be biological interactions since they exist in remotely related homologues. In Figure 6c, the lower the minimum sequence identity of proteins with the same interface in all available CFGs (*N*/*N*), the higher the percentage that is contained in PDB, PQS and PISA. For all values of *N*, when the minimum sequence identity is 60% or less, more than 95% of PISA assemblies contain the interface.

Evidence for previously contested dimer interfaces

Comparison of crystal forms has allowed us to confirm many biological interfaces in homooligomeric 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 previously been considered as part of biologically relevant structures. For 35 families with 2 or 3 crystal forms involving 145 entries, one or more common interfaces are detected in all crystal forms with minimum identity < 50%, which are not annotated or only rarely annotated in the PDB or PQS or PISA. From our benchmarking analysis, those interfaces are likely to be biological relevant. In these cases, it is valuable to have solution experimental data that indicate that a protein is a homooligomer under approximately physiological conditions. We give two examples, shown in Figure 7.

The endosomal sorting complex I required for transport (ESCRT-1) 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 ESCRT-I complex[54]. The VPS28 C-terminal domain-like (a.24.28.1) family contains 3 crystal form groups and 3 entries (2J9W, 2J9V, 2G3K), which share a common interface (Figure 7a) with minimum identity 39% between the yeast and *Xenopus laevis* structures. The full length Vps28 can form homodimers in solution[54], while the C-terminal domain alone is monomeric. However, the common interface in three crystal form groups with sequence identity 39% indicates that the C-terminal domain may also form a homodimer, probably stabilized in the presence of the homodimeric N-terminal

domain. The PDB and PQS do not have the interface in the biological units for any of these 3 entries, while it is in the PISA assemblies for entries 2J9V (yeast Vps28) and 2J9W (*Xenopus* Vps28) but not 2G3K (yeast Vps28). For 2G3K, PISA 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 family *Runt domain* (b.2.5.6), there exists an interface (Figure 7b) formed by two Runt domains with 3 crystal form groups containing 5 entries. The interface comprises a two-stranded parallel β sheet consisting of strands β3 and β12, and part 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[55]. Moreover, one of five buried water molecules W1 is part of a major groove-binding motif that is conserved in Runt and STAT proteins[56]. Only PQS contains this dimer in the biological unit of one of these PDB entries (1EAQ). PDB and PISA do not have the interface in any of these 4 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 exists few repositories of experimental data on the oligomeric state of proteins in solution, such as gel filtration, analytical ultracentrifugation, or 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 oligomerization state, and therefore may fail to identify some physiologically relevant states. Two databases that are available are Doodle and BRENDA. Doodle (http://dimer.tamu.edu/doodle/)[57; 58; 59] 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 *E. coli* and yeast genomes. Another useful database is BRENDA (http://www.brendaenzymes.info)[60], which collects information on enzymes based on enzyme classification (EC) 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. While some proteins are “well known” to be dimers or tetramers, the experimental data on which that is ultimately based are often obscure. This is a common problem in other areas such as gene ontology[61], 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 papers 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 correct but certainly not always[36]. In addition to surface area and symmetry, higher residue conservation in one interface compared to another may indicate those more likely to be biological[62]. Mutation data may also be available that indicate residues likely to be in a biological interface.

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 that for PDB entries a benchmark constructed from the data of Bahadur et al.[36] and Ponstingl et al.[35] that monomeric proteins rarely have common interfaces in more than one third of their crystal forms and usually much less, while multimeric proteins have common interfaces in a majority and frequently all of 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 dataset 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 crosslinking 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.

Methods and materials

Data Sources

The data used in this study come from five sources: protein structure files from the PDB in XML format (PDBML)[19; 63]; biological unit coordinate files from PQS[20] in the legacy PDB format; PISA multimers in XML format[21]; domain classification files from SCOP 1.73[24]; and CE/PSI-BLAST hit files from a non-redundant (100%) PDB database of our lab[64; 65]. We build a crystal form and PDB biological unit from the asymmetric unit information given in the PDB XML files. We are using the most recently remediated XML files released in beta form in April 2007 and in final form on August 1, 2007 from http://www.pdb.org.

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 sequence identity > 50% and alignment length > 80% of both proteins or 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 are combined into crystal from groups 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 highest X-ray resolution as the *representative* entry of the crystal form to compare interfaces between two different crystal forms.

Building crystal structures—A crystal structure is 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 operator were used to build the ASU first. The symmetry operators for forming each PDB biological unit in Cartesian (Ångstrom) form were provided by the PDB (Zukang Feng, personal communication).

Building crystals using space group symmetry operators: The scale matrix is read from the PDB XML file for each entry:

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

The fractional coordinates are therefore:

$$\mathbf{X}_f=\mathbf{SX}$$

where $\mathbf{X} = (x\ y\ z\ 1)^{\mathrm{T}}$ 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

$$\mathbf{P}_m=\begin{pmatrix} p_{11} & p_{12} & p_{13} & t_1 \\ p_{21} & p_{22} & p_{23} & t_2 \\ p_{31} & p_{23} & p_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Here, $(t_1\ t_2\ t_3)^{\mathrm{T}}$ is the translation vector. A fractional coordinate is transformed by $\mathbf{P}_m$ to $\mathbf{X}_m$.

$$\mathbf{X}_m=\mathbf{P}_m\mathbf{X}_f=\mathbf{P}_m\mathbf{SX}$$

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 construct 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 apply 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 translate these symmetry copies into this unit cell by adding 0,±1,±2, etc. as necessary.

To identify contacts among monomers in the crystal, we build 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 results in a 3×3 collection of unit cells. These translations are formed from the scaled unit cell coordinates by adding 1 or −1 to the coordinates:

$$\mathbf{X}_{mn}=\mathbf{X}_m+\mathbf{T}_n$$

where $\mathbf{T}_n = (i\ j\ k\ 0)^{\mathrm{T}}$ and $(i\ j\ k)^{\mathrm{T}}$ is the nth vector in the series: $(-1\ -1\ -1)^{\mathrm{T}}$, $(-1\ -1\ 0)^{\mathrm{T}}$, $(-1\ -1\ +1)^{\mathrm{T}}$,…,$(+1\ +1+1)^{\mathrm{T}}$.

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

$$\mathbf{X}'_{mn}=\mathbf{S}^{-1}\mathbf{X}_{mn}=\mathbf{S}^{-1}(\mathbf{P}_m\mathbf{S}\mathbf{X}+\mathbf{T}_n)$$

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 tree of bounding boxes. The method is based on *k*-dimensional discrete oriented polytopes (*k*-dops)[66]. These are multi-faceted bounding boxes with faces perpendicular to one of *k* fixed axes. We use *k*=3, corresponding to the three Cartesian axes. A *k*-dop 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 yes, it is a leaf node, with a bounding box containing the atoms of the residue. Otherwise, go to step 3.

We build a tree for each chain in the unit cell containing the original asymmetric unit. To save time and computer memory, the same tree is used for all other unit cells. However, because of the translations and rotations, the bounding boxes for these unit cells must be rebuilt, starting from the leaves of the tree and working our way up to the each tree root, that is each protein chain. All atomic contacts are calculated whenever there are overlaps between two leaf nodes.

Comparing interfaces—To compare interfaces between two different crystal forms, we compare 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 Cβ atoms (Cα for glycine) with distance less than 12Å. An atomic contact exists between residues with at least one atomic distance less than 5Å. The interfaces of an entry are considered to be the same (or non-unique) when they consist of the same pair of chains from the asymmetric unit and have *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 paper[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 require 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 Cβ-Cβ distances of corresponding amino acids in the two interfaces (we use Cα 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_{ij} and f_{ij}, and the weights w_{ij} are some monotonically decreasing function of $d_{ij} = \min(e_{ij}, f_{ij})$ then *Q* is defined:

$$Q = \frac{\sum_{i,j} w_{ij} \exp(-0.5\,|e_{ij} - f_{ij}\,|)}{\sum_{i,j} w_{ij}}$$

The similarity function uses a distance-weighted score with a weight of 0 for Cβ-Cβ distance greater than or equal to 12Å (Cα for Gly). This value was selected by calculating the Cβ-Cβ distances for a large set of protein dimers from the PDB and measuring the probability of an atomic contact (≤5 Å) if the Cβ-Cβ distance was less than *D*. As shown in Figure 8(a), this probability, *p* (atom contact $\|\mathbf{r}(C\beta_i) - \mathbf{r}(C\beta_j)| \le D$), 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 $D \ge 5$ Å and 1.0 if $D < 5$ Å, as shown in Figure 8(b). Therefore the weight function is defined:

$$y = \begin{cases} 1 & 0 < x \le 5 \\ 2.08\exp\left(\frac{-(x-5)^2}{5.159}\right) & x > 5 \end{cases}$$

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 non-bonded 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 select a list of homodimers with sequence identity < 20% each other, and calculated the minimum Cβ distances for each Cβ in the interface. The distribution of Cβ distances fits into a Gaussian distribution (Figure 8c).

$$y = 0.036 * e^{\left(-\frac{(x - 17.31)^2}{272.38}\right)}$$

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 (Figure 8d). When Q score > 0.08, the probability for Q score from two arbitrary dimers is close to 0. In our study, we set the minimum Q score to be 0.1

$$y = 0.114 * e^{\left(-\frac{(x - 0.026)^2}{0.00039}\right)}$$

Surface Area

The surface area of each unique interface in each crystal form is calculated with the program NACCESS[68].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 two:

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

Clustering Family Interfaces—We clustered interfaces with surface area greater than 200 Å^2 that occur in at least two crystal form representatives using a hierarchical clustering algorithm. Interfaces 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 Q > 0.20 with at least half of the interfaces already in cluster *C*.

Benchmark Data Set

The benchmark data set was compiled from the Ponstingl[35] and Bahadur[36; 37] data sets in our study to verify if common interfaces in multiple crystal form groups 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 (http://firebird.sourceforge.net/). 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, 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 biological unit interface comparison. All interfaces that occurs at least two crystal forms are output as PDB formatted files.

Availability

Structures of protein dimers observed in multiple crystal forms are available from the authors at http://dunbrack.fccc.edu/JMB08/InterfaceFiles.tar (250 MBytes). Phylogenetic trees of SCOP families with common interfaces are available at http://dunbrack.fccc.edu/JMB08/SupplementaryFigures.tar (2.5 MBytes). Other supplementary material is available from the journal website.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Dr. Longin Jan Latecki for useful comments. We thank Zukang Feng and Eugene Krissinel for providing information required to build PDB and PISA 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. This work was supported by NIH grant R01 GM73784 to RLD.

References

1. Traut TW. Dissociation of enzyme oligomers: a mechanism for allosteric regulation. Crit Rev Biochem Mol Biol 1994;29:125–163. [PubMed: 8026214]
2. Jaenicke R. Stability and folding of domain proteins. Prog Biophys Mol Biol 1999;71:155–241. [PubMed: 10097615]
3. Ispolatov I, Yuryev A, Mazo I, Maslov S. Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 2005;33:3629–3635. [PubMed: 15983135]
4. Yarden Y, Schlessinger J. Epidermal growth factor induces rapid, reversible aggregation of the purified epidermal growth factor receptor. Biochemistry 1987;26:1443–1451. [PubMed: 3494473]
5. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A 1996;93:13–20. [PubMed: 8552589]
6. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 1995;63:31–65. [PubMed: 7746868]

7. Schlunegger MP, Bennett MJ, Eisenberg D. Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. Adv Protein Chem 1997;50:61–122. [PubMed: 9338079]
8. Bennett MJ, Schlunegger MP, Eisenberg D. 3D domain swapping: a mechanism for oligomer assembly. Protein Sci 1995;4:2455–2468. [PubMed: 8580836]
9. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI. Structural similarity enhances interaction propensity of proteins. J Mol Biol 2007;365:1596–1606. [PubMed: 17141268]
10. Gensure RC, Makitie O, Barclay C, Chan C, Depalma SR, Bastepe M, Abuzahra H, Couper R, Mundlos S, Sillence D, Ala Kokko L, Seidman JG, Cole WG, Juppner H. A novel COL1A1 mutation in infantile cortical hyperostosis (Caffey disease) expands the spectrum of collagen-related disorders. J Clin Invest 2005;115:1250–1257. [PubMed: 15864348]
11. Vorgerd M, van der Ven PF, Bruchertseifer V, Lowe T, Kley RA, Schroder R, Lochmuller H, Himmel M, Koehler K, Furst DO, Huebner A. A mutation in the dimerization domain of filamin c causes a novel type of autosomal dominant myofibrillar myopathy. Am J Hum Genet 2005;77:297–304. [PubMed: 15929027]
12. Fridovich I. Superoxide dismutases. Adv Enzymol Relat Areas Mol Biol 1986;58:61–97. [PubMed: 3521218]
13. Borchelt DR, Lee MK, Slunt HS, Guarnieri M, Xu ZS, Wong PC, Brown RH Jr, Price DL, Sisodia SS, Cleveland DW. Superoxide dismutase 1 with mutations linked to familial amyotrophic lateral sclerosis possesses significant activity. Proc Natl Acad Sci U S A 1994;91:8292–8296. [PubMed: 8058797]
14. Furukawa Y, O'Halloran TV. Posttranslational modifications in Cu,Zn-superoxide dismutase and mutations associated with amyotrophic lateral sclerosis. Antioxid Redox Signal 2006;8:847–867. [PubMed: 16771675]
15. Howlett GJ, Minton AP, Rivas G. Analytical ultracentrifugation for the study of protein association and assembly. Curr Opin Chem Biol 2006;10:430–436. [PubMed: 16935549]
16. Winzor DJ. Analytical exclusion chromatography. J Biochem Biophys Methods 2003;56:15–52. [PubMed: 12834967]
17. Petrotchenko EV, Olkhovik VK, Borchers CH. Isotopically coded cleavable cross-linker for studying protein-protein interaction and protein complexes. Mol Cell Proteomics 2005;4:1167–1179. [PubMed: 15901824]
18. Dafforn TR. So how do you know you have a macromolecular complex? Acta Crystallogr D Biol Crystallogr 2007;63:17–25. [PubMed: 17164522]
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28:235–242. [PubMed: 10592235]
20. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361. [PubMed: 9787643]
21. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. 2007in press
22. Levy ED. PiQSi: protein quaternary structure investigation. Structure 2007;15:1364–1367. [PubMed: 17997962]
23. Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics. 2005
24. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP : a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 1995;247:536–540. [PubMed: 7723011]
25. Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, Choi H, Cho Y, Han K, Lee S, Choi H, Lappe M, Holm L, Kim S, Oh D, Bhak J. PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics 2005;21:2541–2543. [PubMed: 15749693]
26. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 2007;35:D580–D589. [PubMed: 17202171]
27. Winter C, Henschel A, Kim WK, Schroeder M. SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res 2006;34:D310–D314. [PubMed: 16381874]

28. Finn RD, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 2005;21:410–412. [PubMed: 15353450]
29. Liu S, Li Q, Lai L. A combinatorial score to distinguish biological and nonbiological protein-protein interfaces. Proteins 2006;64:68–78. [PubMed: 16596649]
30. Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 2006;7:27. [PubMed: 16423290]
31. Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics. 2004
32. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein-protein interfaces. Proteins 2005;60:353–366. [PubMed: 15906321]
33. Jefferson ER, Walsh TP, Barton GJ. Biological units and their effect upon the properties and prediction of protein-protein interactions. J Mol Biol 2006;364:1118–1129. [PubMed: 17049359]
34. Xu Q, Canutescu A, Obradovic Z, Dunbrack RL Jr. ProtBuD: a database of biological unit structures of protein families and superfamilies. Bioinformatics 2006;22:2876–2882. [PubMed: 17018535]
35. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57. [PubMed: 10944393]
36. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004;336:943–955. [PubMed: 15095871]
37. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53:708–719. [PubMed: 14579361]
38. Shoemaker BA, Panchenko AR, Bryant SH. Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein Sci 2006;15:352–361. [PubMed: 16385001]
39. Faber HR, Matthews BW. A mutated T4 lysozyme displays five different crystal conformations. Nature 1990;348:263–266. [PubMed: 2234094]
40. Hoover DM, Chertov O, Lubkowski J. The structure of human beta-defensin- 1: new insights into structural properties of beta-defensins. J Biol Chem 2001;276:39021–39026. [PubMed: 11486002]
41. Pazgier M, Prahl A, Hoover DM, Lubkowski J. Studies of the biological properties of human beta-defensin 1. J Biol Chem 2007;282:1819–1829. [PubMed: 17071614]
42. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9. [PubMed: 12784359]
43. Li H, Raman CS, Martasek P, Masters BS, Poulos TL. Crystallographic studies on endothelial nitric oxide synthase complexed with nitric oxide and mechanism-based inhibitors. Biochemistry 2001;40:5399–5406. [PubMed: 11331003]
44. Rosenfeld RJ, Garcin ED, Panda K, Andersson G, Aberg A, Wallace AV, Morris GM, Olson AJ, Stuehr DJ, Tainer JA, Getzoff ED. Conformational changes in nitric oxide synthases induced by chlorzoxazone and nitroindazoles: crystallographic and computational analyses of inhibitor potency. Biochemistry 2002;41:13915–13925. [PubMed: 12437348]
45. Love RA, Stroud RM. The crystal structure of alpha-bungarotoxin at 2.5 A resolution: relation to solution structure and binding to acetylcholine receptor. Protein Eng 1986;1:37–46. [PubMed: 3507686]
46. Sutcliffe MJ, Dobson CM, Oswald RE. Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. Biochemistry 1992;31:2962–2970. [PubMed: 1550821]
47. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–998. [PubMed: 14499603]
48. Sommerhalter M, Voegtli WC, Perlstein DL, Ge J, Stubbe J, Rosenzweig AC. Structures of the yeast ribonucleotide reductase Rnr2 and Rnr4 homodimers. Biochemistry 2004;43:7736–7742. [PubMed: 15196016]
49. Lindqvist Y, Huang W, Schneider G, Shanklin J. Crystal structure of delta9 stearoyl-acyl carrier protein desaturase from castor seed and its relationship to other di-iron proteins. Embo J 1996;15:4081–4092. [PubMed: 8861937]

50. Moche M, Shanklin J, Ghoshal A, Lindqvist Y. Azide and acetate complexes plus two iron-depleted crystal structures of the di-iron enzyme delta9 stearoyl-acyl carrier protein desaturase. Implications for oxygen activation and catalytic intermediates. J Biol Chem 2003;278:25072–25080. [PubMed: 12704186]
51. Katzmann DJ, Babst M, Emr SD. Ubiquitin-dependent sorting into the multivesicular body pathway requires the function of a conserved endosomal protein sorting complex, ESCRT-I. Cell 2001;106:145–155. [PubMed: 11511343]
52. Rieder SE, Banta LM, Kohrer K, McCaffery JM, Emr SD. Multilamellar endosome-like compartment accumulates in the yeast vps28 vacuolar protein sorting mutant. Mol Biol Cell 1996;7:985–999. [PubMed: 8817003]
53. Gruenberg J, Stenmark H. The biogenesis of multivesicular endosomes. Nat Rev Mol Cell Biol 2004;5:317–323. [PubMed: 15071556]
54. Pineda-Molina E, Belrhali H, Piefer AJ, Akula I, Bates P, Weissenhorn W. The crystal structure of the C-terminal domain of Vps28 reveals a conserved surface required for Vps20 recruitment. Traffic 2006;7:1007–1016. [PubMed: 16749904]
55. Becker S, Groner B, Muller CW. Three-dimensional structure of the Stat3beta homodimer bound to DNA. Nature 1998;394:145–151. [PubMed: 9671298]
56. Bäckström S, Wolf-Watz M, Grundström C, Hard T, Grundstrom T, Sauer UH. The RUNX1 Runt domain at 1.25A resolution: a structural switch and specifically bound chloride ions modulate DNA binding. J Mol Biol 2002;322:259–272. [PubMed: 12217689]
57. Wang SH, Syu WJ, Hu ST. Identification of the homotypic interaction domain of the core protein of dengue virus type 2. J Gen Virol 2004;85:2307–2314. [PubMed: 15269372]
58. Marino-Ramirez L, Minor JL, Reading N, Hu JC. Identification and mapping of self-assembling protein domains encoded by the Escherichia coli K-12 genome by use of lambda repressor fusions. J Bacteriol 2004;186:1311–1319. [PubMed: 14973045]
59. Marino-Ramirez L, Hu JC. Isolation and mapping of self-assembling protein domains encoded by the Saccharomyces cerevisiae genome using lambda repressor fusions. Yeast 2002;19:641–650. [PubMed: 11967834]
60. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Res 2007;35:D511–D514. [PubMed: 17202167]
61. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32(Database issue):D258–D261. [PubMed: 14681407]
62. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001;313:399–416. [PubMed: 11800565]
63. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM. PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics 2005;21:988–992. [PubMed: 15509603]
64. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591. [PubMed: 12912846]
65. Wang G, Dunbrack RL Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33:W94–W98. [PubMed: 15980589]
66. Klosowski JT, Held M, Mitchell JB, Sowizral H, Zikan K. Efficient collision detection using bounding volume hierarchies of k-DOPs. IEEE Trans. Visualization Comp. Graphics 1998;4:21–36.
67. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem 1983;4:187–217.

68. Hubbard, SJ.; Thornton, JM. NACCESS. Department of Biochemistry and Molecular Biology. London: University College London; 1993.

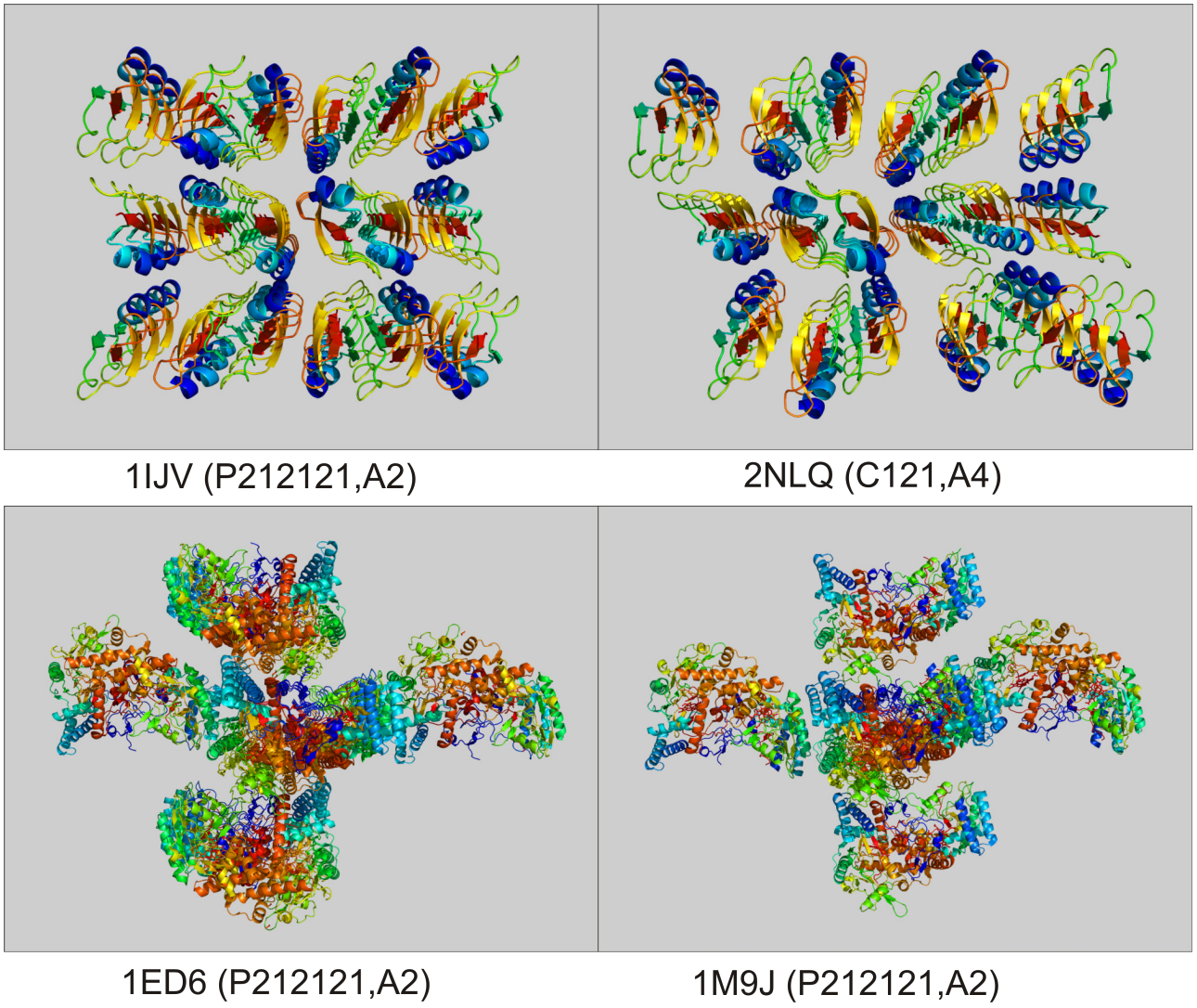

Figure 1.

Examples of similar crystal structures from different space groups. Top: 1IJV with space group P 2_1 2_1 2_1 and asymmetric unit A2, and 2NLQ with space group C 1 2 1 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 $\geq 200 Å^2$. Bottom: 1ED6 and 1M9J with same space group P 2_1 2_1 2_1, same asymmetric unit A2, and quite similar unit cell parameters are not the same crystal form. However, all unique interfaces in 1M9J are contained in 1ED6, while the reverse is not true.

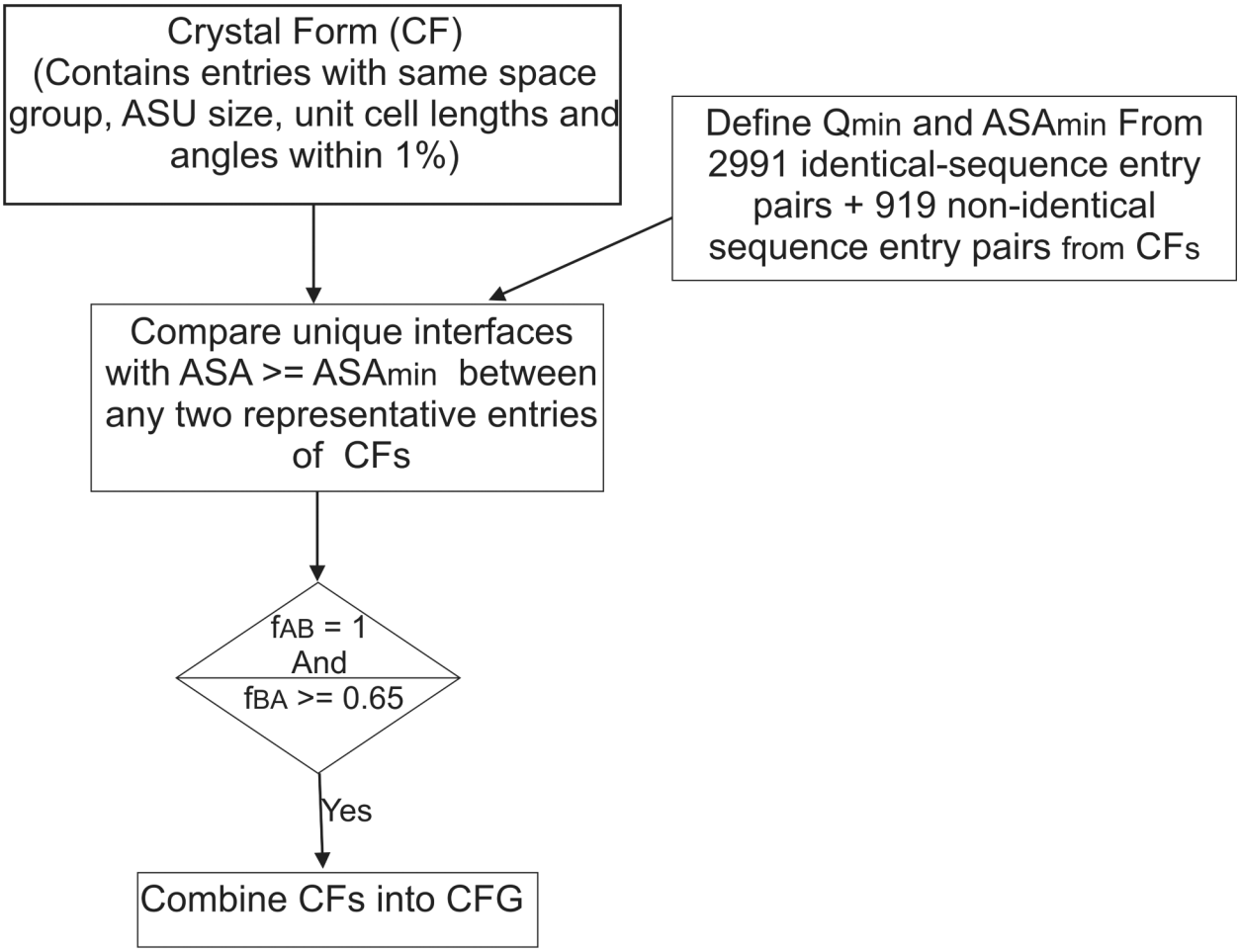

Figure 2.

Flow chart to define a crystal form group (CFG). Q_{min} = 0.1 and ASA_{min} = 200Å^2 are used in this study. f_{AB} = 1 means all interfaces of entry A with ASA ≥ ASA_{min} are similar to some interfaces of entry B with $Q \geq Q_{min}$. $f_{BA} \geq 0.65$ means >~2/3 of interfaces of entry B are similar to some interfaces of entry A. *M* refers to the number of CFGs in an interface cluster, *N* refers to the number of CFGs in a family.

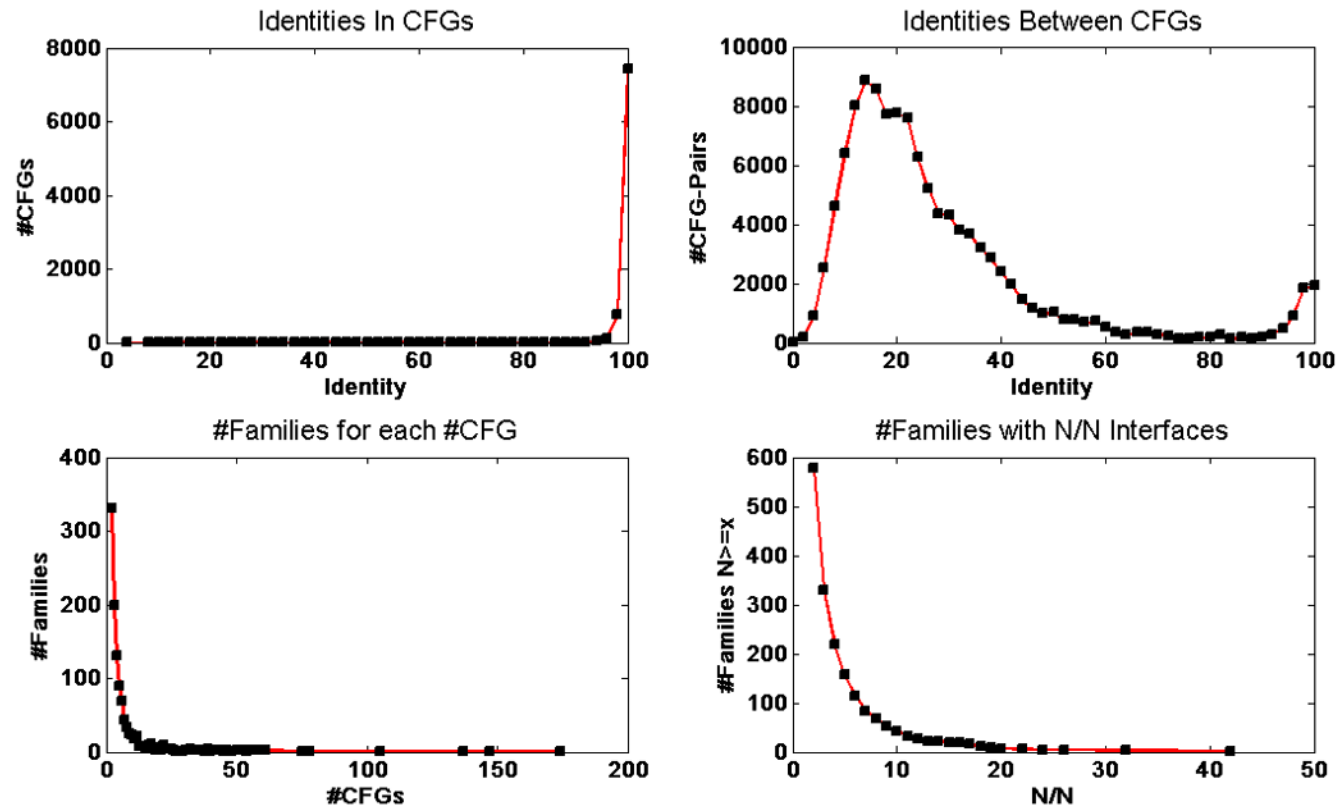

Figure 3.

Overview of families and crystal forms in the database. Top left: Sequence identity distribution in same CFGs. Top right: Sequence identity distribution between different CFGs. Bottom left: Number of families (y-axis) with *N* crystal form groups (x-axis). Bottom right: The relationship between the number of families (y-axis) and the number of crystal form groups (x-axis) such that all *N* crystal form groups for the family contain at least one common interface. The y-axis is the cumulative number of families such that $N \geq x$.

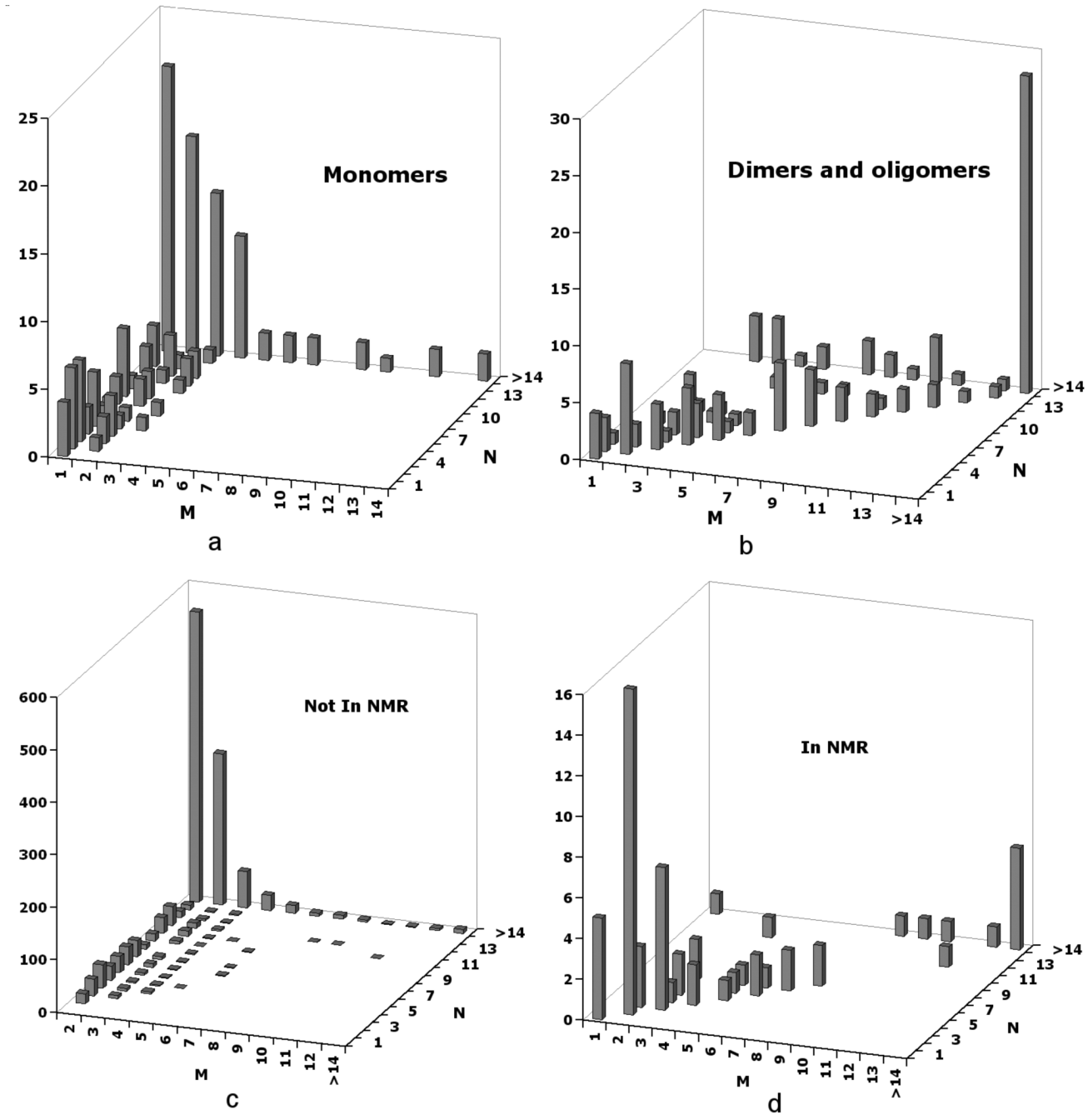

Figure 4.

The correlations between *M*, *N* for 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.

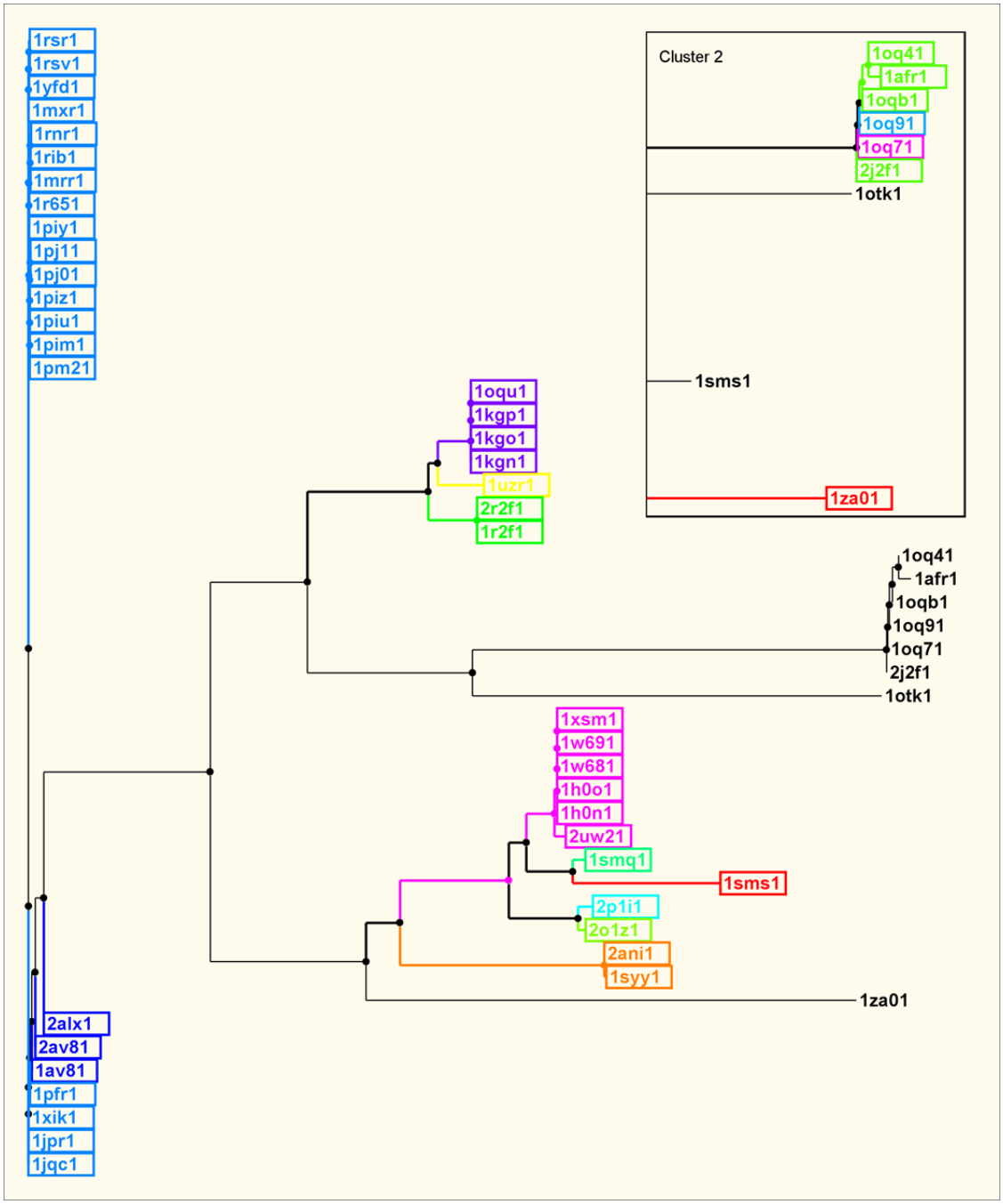

Figure 5.

Phylogenetic tree of *ribonucleotide reductase-like* family (a.25.1.2) with the common interface in 11 of 16 CFGs. Each crystal form group is shown in a different color. Entries without the interface are not boxed and are shown in black type. Inset: different interface present in 4 of 16 CFGs.

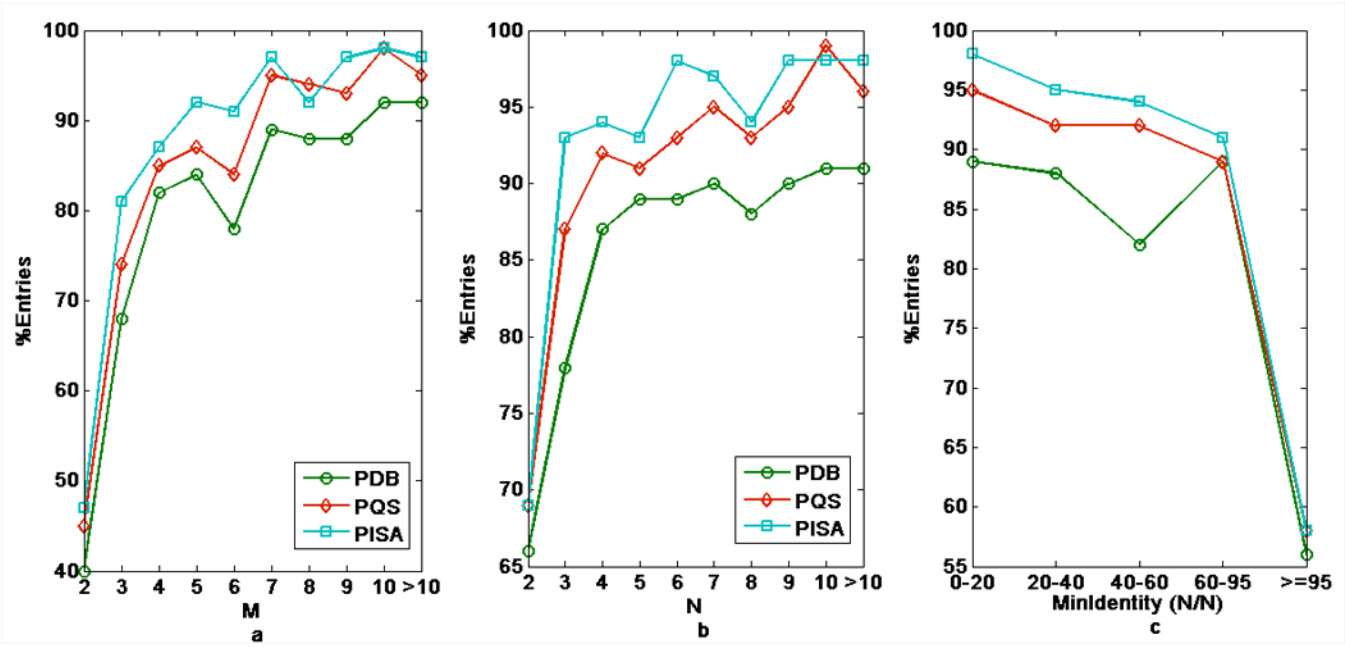

Figure 6.

The percentages (%) of entries in PDB, PQS and PISA that contain common interfaces. (a) % in PDB, PQS and PISA vs. *M* (*M* = 2,3,4,5,6,7,8,9,10,>10) and $M/N \geq 0.5$. (b) % in PDB, PQS and PISA versus *N* (*N* = 2,3,4,5,6,7,8,9,10,>10). (c) % in PDB, PQS and PISA vs. minimum sequence identity ranges for *N*/*N* interfaces.

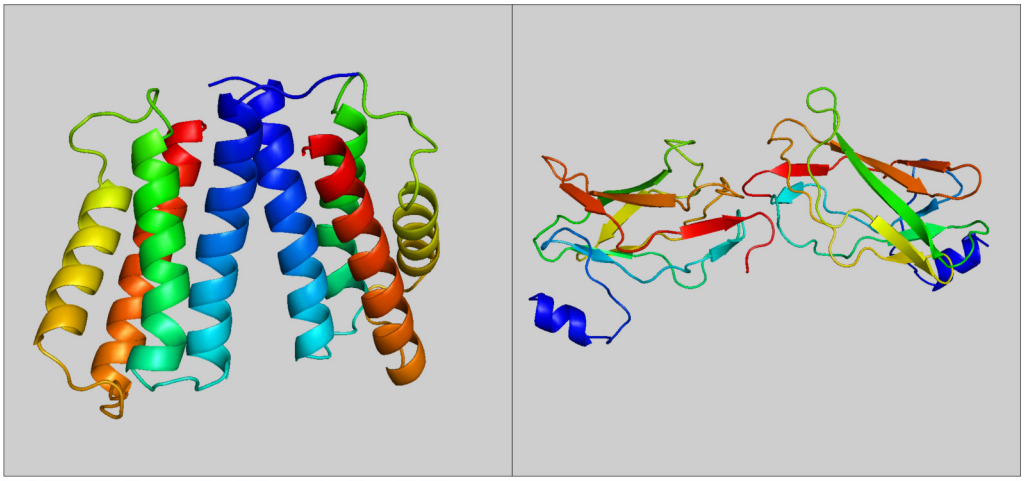

Figure 7.
Examples of interfaces not well annotated in public databases. (a) 2J9W interface (A: X,Y,Z; B: X,Y,Z). (b) 1EAQ interface (A:X,Y,Z; B:1/2-X,1/2+Y,1-Z).

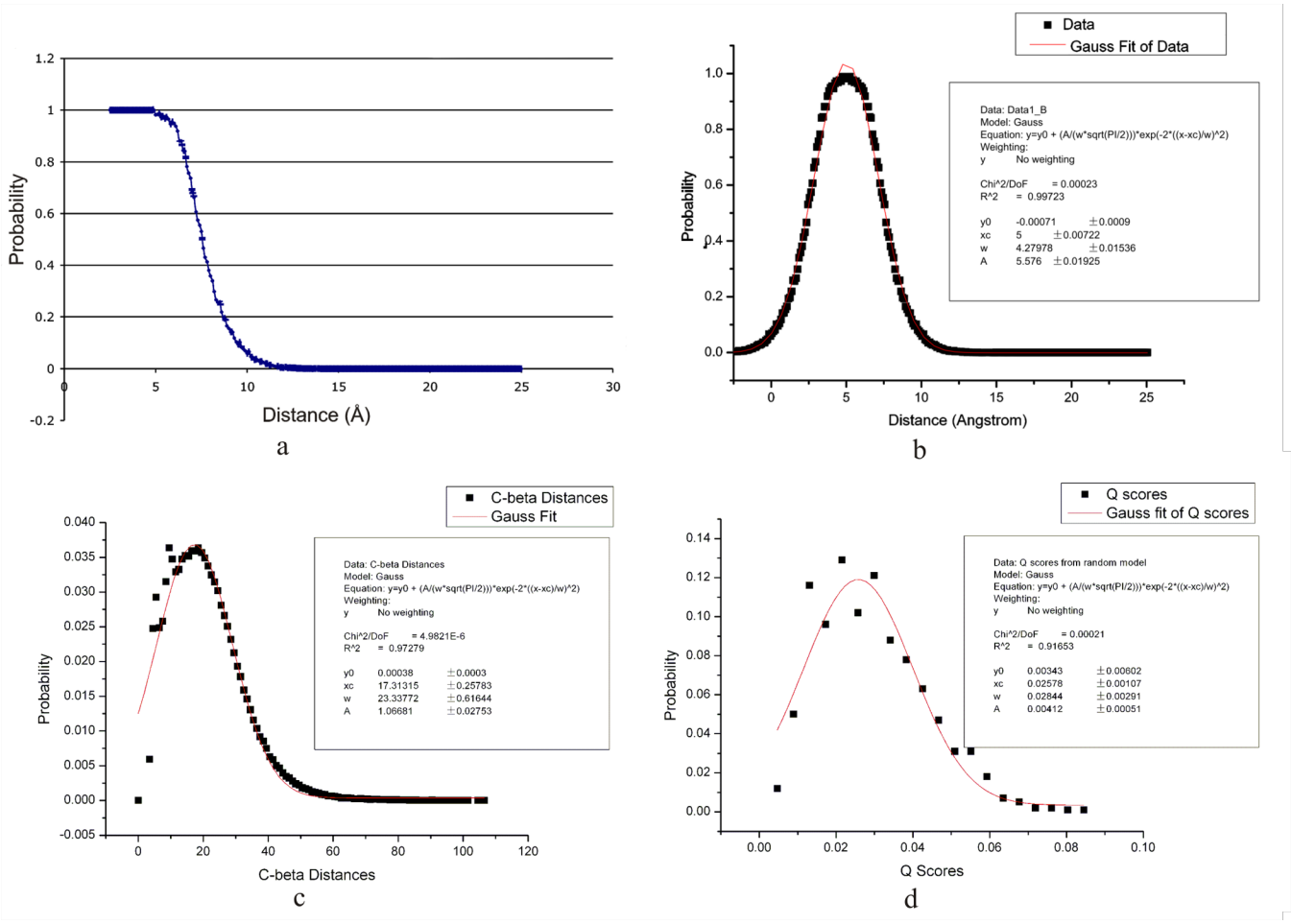

Figure 8.

(a) The relationship between residue distance and contact probability. (b) The Gaussian fit of (a). The data on the left part of b was mirrored from the right part in order to model the data. (c) The distribution of Cβ distances which are calculated from a list of homodimers with sequence identity < 20% each other. (d). The distribution of Q scores computed from 1,000 pairs of distances with size 300 from the Gaussian function in (c). Gaussian fits are modeled by Original Plot 7.0.

Table 1

Numbers of same crystal structures with various minimum surface area and minimum Q scores. Interfaces with ASA ≥ ASA_{cutoff} are in other entry, vice versa. Row: Surface Areas, column: Q scores.

	0	**50**	**100**	**150**	**200**	**250**	**300**	**350**
0	2991	2991	2991	2991	2990	2975	2941	2744
0.05	2567	2907	2984	2987	2986	2971	2941	2744
0.10	2549	2902	2984	2987	2986	2971	2941	2744
0.15	2541	2880	2961	2964	2986	2971	2941	2744
0.20	2539	2879	2960	2963	2985	2970	2941	2744
0.25	2532	2873	2954	2958	2983	2968	2940	2744
0.30	2521	2861	2941	2946	2983	2968	2940	2744
0.35	2507	2843	2924	2929	2983	2968	2940	2744

Table 2

For structure pairs with identity < 100%. Structure pair is between representative entry and the entry with minimum sequence identity in each CF. Interfaces with ASA ≥ ASA_{cutoff} are in other entry, vice versa. Row: Surface Areas, column: Q scores.

	0	**50**	**100**	**150**	**200**	**250**	**300**	**350**
0	919	919	919	919	919	909	899	862
0.05	699	839	891	901	907	902	894	858
0.10	688	834	888	899	905	900	892	856
0.15	684	831	885	897	903	899	891	855
0.20	682	826	878	890	898	895	887	853
0.25	680	823	878	890	896	893	885	851
0.30	678	822	878	890	896	893	885	850
0.35	669	813	869	883	890	888	881	847

Table 3

Overview of common interfaces.

	#Family	#CFGs	#Entry	In PDB (%)**	In PQS (%)	In PISA (%)	In ASU (%)
Single-domain Structures	1377	8816	19842	-	-	-	-
≥2 CFGs/Family	1125	8564	19446	-	-	-	-
≥2 CFGs/Cluster	868	6292	15264	-	-	-	-
N/N (N ≥ 4)	176	1372	3139	90	95	97	58
1>M/N > 0.7 (N ≥ 4)	74	676	1781	88	91	93	61
N/N (N = 2, 3)	266	618	1049	70	76	80	52
M ≥ 4 (M/N ≥ 0.5)*	248	2136	5200	89	93	95	59

* From the analysis of Ponstingl/Bahadur benchmark set and NMR structures, M ≥ 4 and M/N ≥ 0.5 is most likely to be biological interest.

** % in PDB, PQS, PISA and ASU are based on clusters, may have duplicate entries. NMR entries are excluded.

Table 4

Biological units of benchmark entries in public databases

	PDB	PQS	PISA
Benchmark monomers (132)	88 monomers 41 dimers 3 larger oligomers	60 monomers 55 dimers 17 larger oligomers	87 monomers 36 dimers 9 larger oligomers
Benchmark dimers (84)	3 monomers 78 dimers 3 larger oligomers	1 monomers 76 dimers 7 larger oligomers	1 monomers 80 dimers 3 larger oligomers
Benchmark oligomers (size>2) (42)	1 monomers 1 dimers 40 larger oligomers	1 monomers 0 dimers 41 larger oligomers	1 monomers 1 dimers 40 larger oligomers

Table 5

Interfaces with M ≥ 5 from benchmark monomers.

PdbID	N Family	M Cluster	#Entries Family	#Entries Cluster	ASA	Min SeqID	In PDB	In PQS	In PISA	#PDB*	#PQS*	#PISA*
232l	34	14	443	413	731	77	0	1	0	0	324	11
256l	34	14	443	413	762	77	0	1	0	0	324	11
1caq	38	12	78	31	672	58	0	1	1	3	21	10
830c	38	12	78	31	707	58	0	1	1	3	21	10
1kwa	44	10	52	12	579	12	0	1	1	2	3	3
1afk	48	9	175	72	822	63	0	1	1	2	43	46
2abx	19	9	31	13	363	20	1	1	0	9	11	4
1clu	137	7	212	34	754	26	0	1	0	18	24	3
1vlz	50	7	94	25	404	21	0	0	0	0	0	0
1rb3	32	6	118	25	616	31	1	1	0	18	23	2
1mwc	55	6	337	24	299	12	0	0	0	1	1	1
1bin	55	5	337	11	561	16	0	0	1	5	4	4

* number of entries in PDB, PQS and PISA which contain the common interfaces.

Table 6

Probability (P (Biological interface | M, N)) for each M, N and minimum identity ≥ 90, derived from 2,185 monomers and 254 dimers/oligomers in benchmark data set and PIQSI.

	2	3	4	5	6	7	8	9	10	11	12	13	14	> 14
1	0.20	0.15	0.13	0.18	0.07	0.04	0.24	0.08	0.06		0.09	0.50	0.50	0.07
2	0.58	0.40	0.09	0.07	0.03	0.18	0	0.07	0.13	0	0.27	0		0.05
3		0.67	0.25	0	0.60	1			0		0.33	0		0.04
4			1							0	0.33			0
5					0									0.06
6														0
7														1
9														0

*Row: M, column: N. Empty cells mean no data points are available. Red means the total data points ≤ 10.

Table 7

Probability (P (Biological interface | M, N)) for each M, N and minimum identity < 90, derived from 1,986 monomers and 2,836 dimers/oligomers in benchmark data set and PIQSI.

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**>14**
2	0.95	0.08	0.3	0.5	0	0.5	0.15	0	0	0.5	1	0		0.10
3		0.95	1	0.67	0.76	0.88	0.1	1	0.31	0	0.14	0.14	0	0.11
4			1	1		1	1	0.5			1	0		0.34
5				1	1	1	1	1	1	1	1			0.33
6					1	1			0.5		0.67			0.25
7						0.99					1			0.39
8							0.98	1			1			0.45
9								1			1			0.39
10									1	1				0.92
11										0.99				0.88
12											1			0.77
13												1	1	0.75
14													1	1*
>14														0.99

*Exclude 358 lysozyme entries with 14/34 common interface and minimum identity = 77.

Table 8

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

Family Name	SCOP Code	N Family	#Entries Family	PDB BU	PQS BU	PISA BU	ASU	ASA	Min Identity
AAT-like	c.67.1.1	42	152	149	149	152	83	2998	7
GABA-aminotransferase-like	c.67.1.4	42	123	104	119	123	74	3791	6
Cystathionine synthase-like	c.67.1.3	32	51	47	50	51	33	2560	6
Crotonase-like	c.14.1.3	26	35	31	34	34	26	1459	12
PNP-oxidase like	b.45.1.1	24	36	34	35	36	20	1761	6
Decarhsidesylase	c.1.2.3	22	55	42	55	54	35	1815	12
TNF-like	b.22.1.1	20	28	27	27	28	20	982	5
Nucleoside diphosphate kinase, NDK	d.58.6.1	19	59	59	58	58	33	893	25
L-aspartase/fumarase	a.127.1.1	18	34	31	33	34	17	3436	12
		18	34	32	33	34	16	2124	12
		18	34	32	33	34	27	1528	12
dUTPase-like	b.85.4.1	17	48	38	47	48	21	1669	14
Alcohol dehydrogenase-like, C-terminal domain	c.2.1.1	17	20	16	17	20	12	1462	13
Antibiotic resistance proteins	d.32.1.2	17	27	19	18	25	23	1933	8
PaaI/YdiI-like	d.38.1.5	17	29	24	25	27	20	1151	8
YjgF/L-PSP	d.79.1.1	17	20	17	19	19	14	1037	12
Citrate synthase	a.103.1.1	16	27	24	26	26	14	4820	21
Glutaminase/Asparaginase	c.88.1.1	16	29	28	28	29	25	2162	20
dTDP-sugar isomerase	b.82.1.1	15	22	20	21	22	14	1607	27
NADH oxidase/flavin reductase	d.90.1.1	15	30	30	30	30	27	3571	10
		428	893	804	857	884	554		

Table 9

SCOP families with N≥ 16, M ≥ 7 and M/N ≥ 0.7.

mily Name	SCOP Code	N Family	M Cluster	#Entries Family	#Entries Interface	PDB BU	PQS BU	PISA BU	ASU	ASA	Min Identity
rosine-dependent oxidoreductases	c.2.1.2	147	111	338	251	233	247	247	148	1693	4
osephosphate isomerase (TIM)	c.1.1.1	39	33	81	75	73	74	75	65	1587	17
yptophan synthase beta subunit-like PLP-dependent zymes	c.79.1.1	34	26	48	39	35	39	38	31	2099	11
meric isocitrate & isopropylmalate dehydrogenases	c.77.1.1	25	24	68	67	67	67	67	30	2621	12
sitol monophosphatase/fructose-1,6-bisphosphatase-like	e.7.1.1	25	22	97	90	75	88	87	63	1900	10
DH-like	c.82.1.1	23	22	66	65	63	64	65	52	2704	11
ritin	a.25.1.1	23	20	88	85	82	81	85	54	1285	4
		23	20	88	85	82	80	81	56	527	4
		23	17	88	77	74	72	74	46	478	4
ss I aminoacyl-tRNA synthetases (RS), catalytic domain	c.26.1.1	23	18	50	45	37	39	45	16	1669	10
DH-like	c.82.1.1	23	17	66	45	44	43	43	34	1284	11
PS sulfotransferase	c.37.1.5	22	16	34	27	5	216	0	5	374	22
troviral protease (retropepsin)	b.50.1.1	20	19	220	219	215	114	216	199	1765	15
ymidylate synthase/dCMP hydroxymethylase	d.117.1.1	20	19	120	119	115	54	119	49	2346	15
cleoside diphosphate kinase, NDK	d.58.6.1	19	17	60	55	53	14	52	31	799	25
rR-like transcriptional regulators	a.4.5.28	18	15	18	15	14	23	14	8	2521	8
iquitin-related	d.15.1.1	18	13	25	15	7	23	6	10	621	32
utaminase/Asparaginase	c.88.1.1	16	12	29	24	23	1346	24	13	1041	43
		16	12	29	24	23	247	24	16	935	43
		557	453	1613	1422	1320	74	1362	926		