# Biological Units and their Effect upon the Properties and Prediction of Protein-Protein Interactions

*Emily R. Jefferson, Thomas P. Walsh and Geoffrey J. Barton**

University of Dundee, School of Life Sciences, Dow Street, Dundee, DD1 5EH Scotland, UK

Running title – PQS, effect upon protein-protein interactions

*Corresponding Author

**Keywords**: PQS, Protein-protein interaction, Domain-domain interaction, Protein-protein interface analysis, Prediction

## Summary

Structural data as collated in the Protein Data Bank (PDB) have been widely applied in the study and prediction of protein-protein interactions. However, since the basic PDB Entries only contain the contents of the asymmetric unit rather than the biological unit, some key interactions may be missed by analysing only the PDB Entry. 69054 SCOP (Structural Classification of Proteins) domains were systematically examined to identify the number of additional novel interacting domain pairs and interfaces found by considering the biological unit as stored in the PQS (Protein Quaternary Structure) database. The PQS data adds 25965 interacting domain pairs to those seen in the PDB Entries to give a total of 61783 redundant interacting domain pairs. Redundancy filtering at the level of the SCOP family shows PQS to increase the number of novel interacting domain-family pairs by 302 (13.3%) from 2277, but only 16/302 (1.4%) of the interacting domain pairs have the two domains in different SCOP families. This suggests the biological units add little to the elucidation of novel biological interaction networks. However, when the orientation of the domain pairs is considered, the PQS data increases the number of novel domain-domain interfaces observed by 1455(34.5%) to give 5677 non-redundant domain-domain interfaces. 162/1455 novel domain-domain interfaces are between domains from different families, an increase of 8.9% over the PDB Entries. Overall, the PQS biological units provide a rich source of novel domain-domain interfaces that are not seen in the studied PDB Entries and so PQS domain-domain interaction data should be exploited wherever possible in the analysis and prediction of protein-protein interactions.

# Introduction

The availability of the complete genome sequence for more than 200 organisms has led to the identification of many proteins for which there is no known function. Since biological processes often involve protein-protein interactions, knowledge of a protein's interaction partners can provide important clues about function. Accordingly, many experimental and computational techniques have been developed to probe and predict interacting protein partners. High-throughput experimental methods include: yeast two-hybrid[1]; identification by mass spectrometry of isolated protein complexes[2]; and protein chips[3]. Results from these techniques have been made available through several databases[4; 5; 6]. Computational methods have also been devised in an attempt to improve the reliability of results from high-throughput protein-protein interaction experiments[7; 8; 9;].

In parallel with high-throughput experimental approaches, a wide range of computational methods have been developed for protein-protein interaction prediction without 3D structural data. These include: predicting that two proteins will interact if their phylogenetic trees are seen to diverge in a similar manner[10] if correlated mutations between the two proteins are seen in different organisms[11], if the genes of the two proteins are seen in the same order or gene neighbourhood in several organisms[12], or if two domains are seen fused together in one or more organisms but separated in other organisms[13]. Data-mining the literature available[14], machine learning methods based on primary sequence[15] and combinations of the above approaches[16] have also been employed.

Although all the above techniques can yield information on possible interacting partners, the most detailed knowledge regarding individual protein-protein interactions has been

obtained by structural techniques. High-resolution X-ray structures have revealed the subtlety of inter-atomic interactions between two or more protein chains. The distinguishing properties of interaction sites have been characterised to include: residue conservation across species; a tendency to be polar, uncharged and hydrophobic; a planar protruding shape and a higher solvent accessible area[17]. These properties have been exploited to predict interaction surfaces on protein structures[18; 19; 20].

Since three-dimensional data provides the most detailed description of a protein-protein interaction, there has been considerable interest in predicting the structure of complexes from their component structures by docking. Although obtaining a precise atomic-level model of a docked pair of proteins is a difficult problem, the results of recent CAPRI docking prediction challenges suggest progress has been made for complexes where few conformational changes occur on binding[21; 22]. However, significant challenges remain to cope with protein flexibility[23] and to enable accurate large-scale protein-protein docking experiments.

It has been estimated that most protein-protein interactions will conform to one of about 10,000 types[24]. Unfortunately, the current PDB database of known protein three-dimensional structures[25] only represents approximately 2,000 protein-protein interaction types[24]. At the current rate of structure determination it is likely to be another 20 years before there will be a fully representative set of protein-protein interactions available in the PDB[24].

Despite this deficit in data, the more detailed picture of protein-protein interactions that is available from analysis of protein three-dimensional structures has prompted studies to use

observed protein complexes to predict interactions for homologous proteins. The basic principle is that if two proteins A and B are seen to interact in the PDB, then a homologue of A will be predicted to interact with a homologue of B[26; 27]. Detection of more distantly related or analogous protein pairs has been attempted by the development of a multimeric threading method[28]. In contrast, template-based prediction examines interaction site pairs to capture the essential features of the sites and then employs these templates to search for new sites[29].

One problem with working with structures from the PDB is that the coordinates that appear in archive PDB files are those of the asymmetric unit (ASU) which is the fraction of the crystallographic unit cell that has no crystallographic symmetry. This may not be the biologically relevant unit of structure, and so may lack some key protein-protein interactions. A further problem is that some of the interactions seen in the ASU of the crystal may be artefacts of crystallisation and so may not be biologically relevant[30].

The problems of working just from the ASU have led to techniques that apply all relevant symmetry operations and then discriminate between crystal packing artefacts and likely functional protein-protein interactions[31; 32]. The Protein Quaternary Structure (PQS) database[33] is a widely-used database provided by the EBI (European Bioinformatics Institute) which stores probable quaternary structures for proteins that have been identified by applying biologically relevant symmetry operations to the ASU and removing crystal packing artefacts. PQS considers the solvent-accessible surface area, the number of residues that become buried on forming a complex, the difference in solvation energy of folding between the complete Assembly and that calculated for each isolated chain, the number of inter-chain salt bridges and the number of inter-chain disulphide bridges.

Although the paper published in 1998[33] describes PQS as an automatic procedure, the PQS database as available on-line contains complexes that have been screened by expert annotators (K. Henrick personal communication). Screening helps to reduce and correct errors and inconsistencies that may result in a fully automatic procedure.

One alternative to PQS is the RCSB (Research Collaboratory for Structural Bioinformatics) PDB biological units resource. For structures deposited since 1999 the RCSB-PDB provides the coordinates for the complete biological multimeric state of a protein provided by the authors of the structure and the PQS entries for biological multimeric states for proteins deposited before 1999 (Wolfgang Bluhm, personal communication).

The true quaternary state of a complex is not always straightforward to determine and errors are made both by PQS and authors of the structure. PQS was chosen as the data-source for this investigation since although the initial assignments of biological units made by PQS are performed by a computer program, they are hand-curated for each structure and errors and inconsistencies in the PQS database are corrected, and updates made continuously. Thus, the PQS database gets closer and closer to the "accepted" view for all protein biological units. In contrast, the RCSB-PDB biological units are author assigned and so will not change unless updated by the authors themselves. A further advantage of PQS is that it allows for the possibility of multiple complexes for a given protein, whereas the RCSB-PDB only allows one biological state. The PQS data are also available from the EBI in an easy to access SQL format.

Although PQS was chosen instead of the RCSB-PDB biological units analysis of the two sources have shown that 90% of the total assignments are in agreement (Dan Bolser,

personal communication) and therefore it is not thought that using one source of biological units over another will significantly alter the results presented here.

Since it is non-trivial to determine the biological unit for most PDB files without the use of PQS, RCSB-PDB biological units or other such systems, studies of protein-protein interactions have mostly ignored the additional interactions that are potentially available. Accordingly, in this paper the effect of including data from likely biological units as stored in PQS on observed domain-domain interactions has been systematically investigated.

## Results and Discussion

The terminology employed by the MSD database[34; 35] was adopted throughout this study. An ASU seen in the PDB file is called an “Entry”. A likely biological unit generated by PQS[33] is called an “Assembly”. Assemblies represent structures with the relevant symmetry operations applied and without the non-biological contacts seen in the ASU. One or more Assemblies may be derived from a single Entry. Figure 1 illustrates this terminology and summarises the generation of Assemblies from Entries by PQS.

In Figure 1(a) the chains of the Entry do not make contact with each other in three-dimensional space. As a consequence, if this Entry was used for analysis or prediction of protein-protein interactions, the conclusion would be that the chains do not interact. However, when the relevant symmetry operations are applied to create the Assembly, the chains can clearly be seen to interact with each other.

In contrast to Figure 1(a), Figure 1(b) illustrates the effect of removing the crystal packing

artefacts seen in an Entry. PQS considers the interaction between the two chains seen in the Entry to be due to crystal packing rather than a true biological interaction and so divides the structure into two different Assemblies. Finally, Figure 1(c) illustrates how an Assembly may result both from removal of crystal packing artefacts and application of the relevant symmetry operation.

## *Extra Interactions Generated by Applying Likely Biological Symmetry Operations*

Protein structural domains rather than chains provide a more accurate basis for classifying functional groups since there can be more than one functional group in a single chain. Accordingly, the analysis presented here was based entirely on interactions seen between domains defined by the hierarchical SCOP structural domain classification[36] rather than protein chains.

The differences between the domain interactions revealed by symmetry operations and those already present in the Entry coordinates were compared by examining domain-domain interactions seen in the same Assembly as illustrated in Figure 2. The domains in each Assembly which were also present in the Entry were classified as “PDB domains”. Domains generated by applying symmetry, and so only present in the Assembly, were classified “PQS domains”. Domain-domain interactions were then classified as “PDB interactions” (interactions between domains that were both present in the Entry) or “PQS interactions” (interactions where at least one of the interacting domains was generated by symmetry). As a consequence of working from PQS Assemblies, any PDB interactions which PQS predicted to be crystal packing artefacts were excluded from the data set taken

forward for analysis.

The first column in Table 1 summarises the interactions seen across all of the Assemblies. There were 35818 interacting domain pairs in the PDB set and a further 25965 interacting pairs from PQS. Thus PQS provides 72.5% more domain-domain interactions than are available in the PDB ASU. However, this data set is redundant since it contains many interactions which are similar to each other and could be considered equivalent. In order to determine the number of additional non-redundant interactions from PQS, a set of interacting domain pairs from PQS which had no equivalent PDB interaction was created.

The hierarchical classification of SCOP[36] was applied to compare domain equivalence as discussed in the methods section. Given the classification into superfamily and family sets, the number of PQS and PDB interacting pairs in each set was determined. If a set of interactions for a particular non-redundant superfamily or family only contained PQS interactions, it was considered a non-redundant PQS interaction.

The three columns labelled “Non-redundant – family level similarity” in Table 1 show the spread of different interaction types at the family level of similarity. The total number of non-redundant interacting domains classified to a family level of equivalence was 2579, of which 302 were seen only in the PQS predicted Assembly. This represents a 13.3% increase in the number of non-redundant domain interactions over those found in the PDB Entries. Most of the additional non-redundant interactions from PQS were interactions between two domains from the same family. In this paper, this type of interaction is termed a “homo-fam-pair” and interactions between two domains from different families a “hetero-fam-pair”. Only 16(1.4%) of the additional non-redundant interactions were hetero-fam-

pairs.

Since as discussed in the introduction, there can be disagreement between the biological unit seen in PQS, the RCSB-PDB and the authors of the structure, the 16 additional hetero-fam-pair non-redundant interactions generated by PQS were investigated in detail by comparison to the literature and to the RCSB-PDB biological units database. 10 of the non-redundant interactions showed agreement between the PQS-predicted biologically relevant interaction and the opinion of the structure authors as discussed in the literature. Although 6 of the complexes were not in agreement with the authors of the structure, they are not necessarily in error. The relevant biochemical experiments are not always performed to determine the biologically relevant structure, and so the author's evidence for the quaternary state may not be reliable.

3 of the 10 additional non-redundant hetero-fam-pair interactions generated by PQS in agreement with the authors of the structure were found to come from the Bc1 complex. The Cytochrome $bc_1$ complex (ubiquinol:ferricytochrome *c* oxidoreductase) is found in mitochondria, photosynthetic bacteria and other prokaryotes. The general function of the complex is electron transfer between two mobile redox carriers, ubiquinol ($QH_2$) and cytochrome *c* (cyt *c*). The Entry of Bc1 complex is a monomer of 11 domains. The Assembly shows two-fold symmetry about an axis perpendicular to the membrane plane generating a dimer of 22 subunits. Figure 3(a) shows the Entry and corresponding Assembly with interacting domains highlighted for the interaction between Cytochrome bc1 domain(a.3.1.3) and Rieske iron-sulfur protein (b.33.1.1) from Entry 1be3[37]. This example is described as a dimer in the paper discussing the structure and by PQS but is a monomer according to the RCSB-PDB biological units database. Figure 3(b) shows another

interaction observed in the Bc1 complex, the interaction between Ubiquinol-cytochrome c reductase (f.21.1.2) and Rieske iron-sulfur protein (b.33.1.1) in Entry 1l0n[38]. Again, this example is described as a dimer in the paper discussing the structure and by PQS but is a monomer according to the RCSB-PDB since the PDB file biological unit remark describes the structure as the same as the ASU.

The 10 additional hetero-fam-pair non-redundant PQS interactions which showed agreement with the authors were all observed when two large multi-domain complexes joined together to form a larger complex. This could suggest that these additional interactions are only secondary interactions that would not be observed in isolation from the larger complex. However, the average size of the interaction site is large, involving 18.6 residue pairs (with a range of 10-42 interacting residues) suggesting that many of these domain interactions may also be observed independently of the complex.

An example of an additional non-redundant hetero-fam-pair PQS-generated interaction where the Assembly is not biologically relevant and does not agree with the authors of the structure nor the RCSB-PDB is the structure of the *Paracoccus Denitrificans* four subunit Cytochrome *c* Oxidase complexed with an antibody fragment seen in structure 1qle[39]. PQS predicts two possible Assemblies, one is hexameric (the same structure as the Entry) and the other is 24meric. The additional PQS generated domain-domain interaction is between a Cytochrome *c* oxidase subunit III-like domain (f.25.1.1) and V set domains (antibody variable domain-like) (b.1.1.1). It is interesting to note that although the 24meric complex is unlikely to form *in vivo*, PQS suggests two possible structures. The hexameric structure which is observed in the Entry shows a loss of 3086.1 Ångstroms$^2$ of solvent accessible surface area upon complex formation and shows a gain in solvation free energies of folding

from isolated chains to the complex of -153.45 kcal/mol. The 24meric structure shows values of 3634.8 Ångstroms$^2$ and -731.41 kcal/mol respectively. Since the 24meric Assembly is energetically favourable, PQS suggests the 24meric structure as a biologically relevant structure; however, since the Assembly is mediated by antibody fragments, it is unlikely to be seen *in vivo*. Although the interactions are probably not biologically relevant, they illustrate a possible interaction mode for the domains present in the structure and so might prove valuable in guiding the design of novel interacting molecules.

## *Analysis of different interfaces*

An interacting pair with the same pair-wise family classification as another interacting pair, may interact in different orientations and hence with different interfaces[40;41]. In order to determine how many non-redundant interaction interfaces were observed, the relative orientation of the interacting pair was investigated using an implementation of the iRMSD (interaction root-mean-square deviation) method described by Aloy *et al*[40] and summarised in the methods section.

As shown in Table 1, when orientation was considered, the number of non-redundant family-level interacting pairs increased by a factor of 2.2, from 2579 to 5677. The 5677 pairs that considered orientation were made up of 4222 PDB interactions and 1455 PQS interactions. Thus, the use of PQS accounted for an increase of 34.5% (1455/4222) in non-redundant domain-domain interfaces. The 5677 interacting pairs were broken down further into 3652 homo-fam-pairs and 2025 hetero-fam-pairs, so 64.0% (3652/5677) of the interacting pairs were homo-fam-pairs. In contrast, when orientation was ignored, only 54% (1402/2579) of the interacting pairs were homo-fam-pairs. This difference in ratio between

homo/hetero-fam pairs with and without orientation, suggests that homo-fam-pairs show a higher variability in orientation when compared to hetero-fam-pairs. The average number of orientations for a homo-fam-pair was 2.6 in comparison to 1.7 for the hetero-fam-pairs.

Some families were found to be more highly variable in orientation (have many different interfaces with which they interact) than others. The interactions which were most variable in orientation were all homo-fam-pairs. The trypsin-like serine protease (b.47.1.2) homo-fam-pair had 38 different orientations, of which 19 were PQS generated non-redundant orientations. The pepsin-like protease (b.50.1.2) homo-fam-pair was also highly varied in orientation with 15 PQS and 8 PDB non-redundant orientations. As might be expected, immunologlobin type interactions were highly variable in orientation. V set domain (antibody variable domain-like) (b.1.1.1) homo-fam-pair, I set domain (immunoglobin)(b.1.1.4) homo-fam-pair and C1 set domain (antibody constant domain-like) (b.1.1.2) homo-fam-pair displayed 32 (of which 17 were PQS interactions), 35 (of which 15 were PQS interactions) and 28 (of which 15 were PQS interactions) different orientations, respectively. 33 non-redundant orientations were observed for the protein kinase (catalytic subunit) (d.144.1.7) homo-fam-pair of which 15 were PQS interactions.

The largest number of PQS orientations for any hetero-fam-pair was 3. Hetero-fam-pairs with 3 non-redundant PQS orientations included MHC antigen-recognition domains (d.19.1.1) interacting with C1 set domains (antibody constant domain-like) (b.1.1.2) (11 PDB non-redundant orientations), and amino acid dehydrogenases (c.58.1.1) interacting with amino acid dehydrogenase-like, C-terminal domains (c.2.1.7) (8 PDB non-redundant orientations).

Some of the homo-fam-pairs studied had a high number of non-redundant PQS orientations compared to PDB orientations. For example, the Prokaryotic protease(b.47.1.1) homo-fam-pair showed 6 PQS orientations and 1 PDB. A superposition of all of the orientations observed for this homo-fam-pair is shown in Figure 4. Three of the additional PQS generated orientations for the Prokaryotic protease(b.47.1.1) homo-fam-pair were observed in variants of Htr heat shock proteases. The Htr proteases act as molecular chaperones and monitor the folded state of other proteins. Htr chaperones have a cage structure where the protein to be chaperoned enters and is protected from the cellular environment so that it can fold correctly. The Entries of the structures which have been solved do not show the whole cage structure and the relevant symmetry operations need to be applied to generate the biologically relevant structure.

For each of the non-redundant PQS orientations from the highly variable homo-fam-pairs, the biological relevance was investigated by examination of the literature. For some pairs, the interaction observed in the PQS structure appears necessary to the biological function of the protein. For example, the homo-fam-pair interaction between two Eukaryotic protease (b.47.1.2) domains observed in the structure of the cell death inducing serine protease, Granzyme A (1orf)[58]. Bell *et al*[58] and PQS both agree that Granzyme A has a disulphide-linked quaternary structure. The oligomeric state contributes to substrate selection by limiting access to the active site for potential macromolecular substrates and inhibitors and to substrate specificity in a non-redundant manner by extending the active site cleft.

Some of the PQS structures were found to disagree with the authors of the structure. For example, the crystal structure of the myristylated catalytic subunit of cAMP-dependent

protein kinase (1cmk)[43] is shown by PQS to be a dodecameric structure with multiple homo-fam-pair protein kinase, catalytic subunit (d.144.1.7) interactions, but the authors of the structure consider the biological unit to be that of the Entry (a dimer).

A higher ratio of agreement between the authors of the structures and the PQS predicted biological unit was observed for the less variable interaction orientations. This is thought to be due to the difficulty in distinguishing biologically relevant structures when the energetics of the association appear favourable. As many of the highly orientation-variable non-redundant homo-fam-pairs contain different valid PDB orientations as well as PQS orientations, the PQS system finds it difficult to distinguish the biologically relevant structure.

The fraction of extra PQS interactions was also determined for the superfamily level of similarity (Table 2). At the level of superfamily, without considering orientation, there were 685 (2579 -1894) less non-redundant interactions than at the family level of similarity. As there were a smaller number of non-redundant interactions there were only 150 (8.6%) additional PQS interactions in-comparison to 302 (13.3%) for the family level of similarity. 135 (90.0%) of the additional PQS interactions were in the homo-supfam-pair classification.

There were 5468 different orientations at the superfamily level of similarity, which was only 209 less than at the family level of similarity. In other words, only 209 unique families have the same orientation as another unique family within the same superfamily and are therefore grouped together. Therefore, 96.3% of the time orientation is not conserved between members of the same superfamily pair but different family pairs.

The differences between the PQS and PDB interfaces were also investigated. In summary, PQS interfaces were on average smaller (36.2 residue pairs) than those in the PDB Entry (39.4 residue pairs) possibly due to a bias by crystallographers towards selecting compact asymmetric units. An in-depth analysis of the differences of the interfaces between PQS and PDB interactions will be presented elsewhere.

### *Examination of homodimers and heterodimers determined at the sequence level*

In a complementary analysis, the interacting pairs were classified into homo- or heterodimers based on their sequence similarity to each other, rather than their position in the SCOP hierarchy. Any interacting pair that shared <90% sequence identity after alignment by BLAST[44] was treated as a heterodimer and the rest as homodimers.

The results, summarised in Table 2 for classification by sequence similarity, show that there were 7246 non-redundant interacting pairs observed and 9635 when orientation was considered, an increase of 32.9%. For heterodimers, the increase when orientation was considered, was 25.1%, while for homodimers, it was 42.6%. Thus, homodimers show a larger diversity of interaction orientations than do heterodimers, even when the stringent 90% identity cutoff is employed.

Examining the effect of including PQS interactions shows there were 2183 (29.3%) additional interactions found by PQS when orientation was considered and 1150 (18.9%) when it was not. These increases may be broken down into an increase of 341 (7.3%) PQS interactions for heterodimers and an additional 1842 (65.9%) for homodimers.

*Removal of interactions predicted to be crystal packing artefacts by PQS*

The domain interactions that PQS predicted to be crystal packing artefacts and those predicted to be biologically significant can be distinguished by comparing interactions seen in an Entry to those in the corresponding Assembly or Assemblies. Interactions that are crystal packing artefacts appear in the Entry but not in the Assemblies. Interactions that are biologically meaningful appear in the Assemblies and may also appear in the Entry. For the purposes of this discussion the domain interactions which are seen in the ASU, but not observed in any Assemblies of a particular Entry are called 'PQS-disallowed interactions' and those seen in an Assembly are called 'PQS-allowed interactions'. The set of PQS-allowed interactions also contains those generated by symmetry operations.

There were 1996 PQS-disallowed interactions observed. These PQS-disallowed interactions were classified at the SCOP family level, resulting in 604 non-redundant family pairs. In addition to the 1996 PQS-disallowed interactions there were an additional 985 interactions which were observed in Assemblies, but which PQS marked as due to crystal packing. PQS included these interactions in Assemblies since they are possibly of biological interest. Therefore, there are in total 2981 interactions which PQS classifies as generated by crystal packing. The numbers of PQS allowed interactions seen in Table 1 and Table 2 show that there are 61783 redundant interactions and 2579 non-redundant interactions (at the family level of similarity) that are considered to be biologically relevant interactions by PQS. Accordingly, there are approximately 21 times more interactions in the PDB that are considered to be biologically relevant than are due to crystal packing. This would appear to contradict the claim that most interactions seen in structural data are due

to crystal packing artefacts[30]. This discrepancy may be due to the dataset used in the present study containing only interaction sites of 10 residues or more whereas Valdar *et al*[30] studied interaction sites below this threshold. Interaction sites below 10 residues in size are more likely to be due to crystal packing artefacts.

### *Consistency of PQS*

The consistency of PQS in classifying a particular domain pair as either a true biological interaction or a crystal packing artefact in related Entries was examined. 1386 interactions with the same orientation and family classification were observed in both the PQS-allowed and PQS-disallowed sets. This represents 69.4% (1386/1996) of the crystal packing interactions removed from the data set. However, many of these interactions have been classified correctly in the opinion of the authors of the structures. This suggests that observing a pair of interactions that are equivalent at the family level, with the same orientation, does not guarantee that both will be classified in the same way, PQS-allowed or PQS-disallowed.

## Conclusions

The effect of including data from likely biological units, as stored in PQS, on observed domain-domain interactions has been systematically investigated. The general conclusions are:

1. 302 (13.3%) additional non-redundant SCOP family pair interactions were observed in the PQS biological units.
2. Only 16(1.4%) additional hetero-fam-pair interactions (between members of different

SCOP families) were observed. Therefore, PQS was not found to aid significantly in identifying novel protein-protein interactions except for new homo-fam-pair interactions.

3. 1455 (34.5%) additional non-redundant interaction interfaces were observed in the PQS biological units when the orientation of the domains involved in the interaction was considered.

4. Of the 1455 additional non-redundant interaction interfaces, 164(8.9%) were hetero-fam-pairs.

5. PQS classified 2981 interactions as due to crystal packing artefacts. There were approximately 21 times more interactions that were considered by PQS to be biologically relevant than were classified as due to crystal packing.

## Materials and Methods

This study builds on the MSD data warehouse developed at the European Bioinformatics Institute (EBI)[34; 35]. The MSD warehouse combines the data from the PDB with substantial derived information essential to this study. Derived data includes the likely biological units by PQS[33] and domain definitions by SCOP[36], CATH[45] and Pfam[46].

The MSD data warehouse is currently distributed as a 250GB Oracle relational database. Although the MSD group provide a variety of sophisticated web-based query interfaces to enable straightforward access for general queries (http://www.ebi.ac.uk/msd/index.html) the present study was performed on a locally installed copy of the database to give higher performance and direct access *via* SQL. In order to increase performance further and to allow complex analyses with a high degree of abstraction, the most data intensive parts of the MSD warehouse were migrated to an object-orientated database developed with the Sun Java Data Objects technology (JDO) (FastObjects community edition implementation,

http://www.versant.net/index.html). JDO is an object persistence framework for the Java language which allows the storage, retrieval and querying of objects. The object-oriented database contains a subset of the data from the MSD which needs to be accessed frequently and efficiently together with a mapping to the full MSD data warehouse. This solution provides both efficient access to the whole breadth of the MSD data and fast access to the most frequently queried data. A detailed description of our JDO-based database (termed SNAPPI-DB to stand for Structures, iNterfaces and Alignments for Protein-Protein Interactions – DataBase) will be presented elsewhere (Jefferson, Walsh and Barton, 2006, *manuscript submitted*).

SCOP (version 1.65)[36] domains were assigned to chains for each structure in the database by applying the domain descriptions employed in the MSD[34; 35]. In SCOP, domains are classified in a hierarchy of classes, folds, superfamilies and families. SCOP domains are straightforward to compare since each domain is given a number for each hierarchical level (except the class classification which is given a letter). The combination of these numbers (separated by full stops) gives the whole hierarchical description. Therefore, equality between one domain and another can be determined to different levels of similarity. For example, a single SCOP domain classified 'a.1.2.3' could be classified at four different levels: the class list 'a' for the class level of the hierarchy, the fold list 'a.1' for the fold level, the superfamily list 'a.1.2' for the superfamily level and finally the family list 'a.1.2.3' for the family level of similarity. To determine if one interacting pair was equivalent to another interacting pair at each level of the SCOP hierarchy, equivalence between each of the members of the pairs was found. For example, a SCOP domain pair, 'a.1.2.3' interacting with 'b.1.2.7' when compared to another SCOP domain pair, 'a.1.2.3' interacting with 'b.1.2.4' would be considered to be equivalent at the superfamily, fold and class level but

not at the family level since the last digit in the second partner is different for the two interactions. The interacting pair was treated symmetrically so that a domain pair of 'a.1.2.3' and 'b.1.2.7' was equal to a domain pair of 'b.1.2.7' and 'a.1.2.3'.

Interactions between domains were determined based on distance. Atoms were considered to interact if the distance between them was less than the sum of their van der Waals radii[47] +0.5Å. Two domains were considered to be interacting if there were ≥ 10 interacting residue pairs between the domains. The threshold of 10 residues was chosen based on inspection of interaction sites and study of relevant literature. For example, it has been suggested that most interaction sites which are <20 residues in size are due to crystal packing artefacts[30] while Bolser *et al.*[48] and Park *et al.*[49] found that "the number of domains identified as contacting each other hardly changed for thresholds between one and ten contacts" and chose a threshold of 5 interacting residues. Finally, Aloy *et al.*[40] chose a threshold of 10 residues with an 8Å distance between interacting residues.

In order to determine how many non-redundant interaction surfaces were observed, the relative orientation of the interacting pair was investigated using an implementation of the iRMSD (interaction root-mean-square deviation) method described by Aloy *et al*[40]. This method applies STAMP (Structural Alignment of Multiple Proteins)[50] to align each separate partner of the interaction pairs to be compared and determines the transform of one structure to another. Comparing one pair AB to another pair A'B', the transform of A' on to A and B' on to B is calculated by STAMP. The transform of A' to A is then used to transform A' to A and B' to B. Then for each domain of each pair, 7 representative co-ordinates are selected using the centre of gravity as the middle point and then +/- 5Å in each axis from the centre of gravity to generate 6 further points. The RMSD of the 7 representative points

of A' (7A') on to the 7 points of A (7A) and 7B' onto 7B is then calculated. The transform of B' on to B is then also used to transform A' on to A and the RMSDs of 7A' on to 7A and 7B' onto 7B calculated. The iRMSD for the A transform is the highest RMSD of the 7A' to 7A and 7B' to 7B using the A transform. The iRMSD for the B transform is the highest RMSD of the A domains using the B transform which is going to be the highest RMSD of the 7A' to 7A and 7B' to 7B using the B transform. The overall iRMSD is the lowest of the iRMSD using the A transform or the iRMSD using the B transform.

Aloy et al.[40] suggested that interacting pairs with an iRMSD ≤5Å should be considered similar whereas an iRMSD value as high as 5-10Å could indicate similar positioning of domains but with a rotation of one domain relative to another. In the current work, the data were analysed by inspection of interacting pairs and found to be in agreement with those of Aloy *et al.*[40]. Accordingly, an iRMSD cut-off of ≤5Å was applied to distinguish interactions between pairs that have a similar orientation and those that do not.

Although the iRMSD is an elegant solution to comparing interacting pairs, one possible problem is that large inserts or linkers which are unique to one of the entries may lead to differences in the centre of mass which in turn could cause mis-classification of two interactions into different orientations. In order to determine whether this possibility had a significant effect on the results, the orientations for the family level of similarity were re-run by calculating iRMSD from the center of gravity for the residues where both of the domains overlapped. Excluding linkers in this way did not alter the results significantly and so all data presented in this paper follow the original Aloy and Russell[40] iRMSD calculation.

## Acknowledgements

We thank the MSD group at EBI for discussions and information, Dr Charlie Bond and Prof Daan Van Aalten for advice and discussions regarding crystallography, and Dr Jonathan Monk and Mr Eduardo Damato for network and systems support. We thank Drs Dan Bolser and Kim Henrick for discussions about PQS and the difference between PQS and RCSB-PDB biological units. Emily Jefferson is supported by a BBSRC (UK Biotechnology and Biological Sciences Research Council) studentship and Thomas Walsh was funded by TEMBLOR, European Community Contract No. QLRI-CT-2001-00015.

# References


1. Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6.
2. Anne-Claude, G., Markus, B., Roland, K., Paola, G., Martina, M., Andreas, B., Jörg, S., Jens, M. R., Anne-Marie, M., Cristina-Maria, C., Marita, R., Christian, H., Malgorzata, S., Miro, B., Heinz, R., Alejandro, M., Karin, K., Manuela, H., David, D., Tatjana, R., Volker, G., Angela, B., Sonja, B., Bettina, H., Christina, L., Marie-Anne, H., Richard, R. C., Angela, E., Erich, Q., Vladimir, R., Gerard, D., Manfred, R., Tewis, B., Peer, B., Bertrand, S., Bernhard, K., Gitte, N. & Giulio, S.-F. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. In *Nature*, Vol. 415, pp. 141-7.
3. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M. & Snyder, M. (2001). Global analysis of protein activities using proteome chips. *Science* **293**, 2101-5.
4. Bader, G. D. & Hogue, C. W. (2000). BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465-77.
5. Ioannis, X., Lukasz, S., Xiaoqun Joyce, D., Patrick, H., Sul-Min, K. & David, E. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. In *Nucleic Acids Res*, Vol. 30, pp. 303-5.
6. Andreas, Z., Luisa, M.-P., Michele, Q., Gabriele, A., Manuela, H.-C. & Gianni, C. (2002). MINT: a Molecular INTeraction database. In *FEBS Lett*, Vol. 513, pp. 135-40.
7. Wan Kyu, K., Jong, P. & Jung Keun, S. (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. In *Genome Inform Ser Workshop Genome Inform*, Vol. 13, pp. 42-50.
8. See-Kiong, N., Zhuo, Z. & Soon-Heng, T. (2003). Integrative approach for computationally inferring protein domain interactions. In *Bioinformatics*, Vol. 19, pp. 923-9.
9. Nye, T. M., Berzuini, C., Gilks, W. R., Babu, M. M. & Teichmann, S. A. (2005). Statistical analysis of domains in interacting protein pairs. *Bioinformatics* **21**, 993-1001.
10. Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**, 609-14.
11. Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219-27.
12. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-901.
13. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.
14. Marcotte, E. M., Xenarios, I. & Eisenberg, D. (2001). Mining literature for protein-protein interactions. *Bioinformatics* **17**, 359-63.
15. Gomez, S. M., Noble, W. S. & Rzhetsky, A. (2003). Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* **19**, 1875-81.
16. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. & Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258-61.
17. Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions.

*Embo J* **22**, 3486-92.
18. Jones, S. & Thornton, J. M. (1997). Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **272**, 133-43.
19. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* **307**, 1487-502.
20. Koike, A. & Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel* **17**, 165-73.
21. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I. & Wodak, S. J. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2-9.
22. Mendez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **60**, 150-69.
23. Bonvin, A. M. (2006). Flexible protein-protein docking. *Curr Opin Struct Biol*. **16,** 194-200.
24. Aloy, P. & Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nat Biotechnol* **22**, 1317-21.
25. Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E. & Berman, H. M. (2002). The Protein Data Bank: unifying the archive. *Nucleic Acids Res* **30**, 245-8.
26. Aloy, P. & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* **99**, 5896-901.
27. Pieper, U., Eswar, N., Stuart, A. C., Ilyin, V. A. & Sali, A. (2002). MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res* **30**, 255-9.
28. Lu, L., Lu, H. & Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350-64.
29. Aytuna, A. S., Gursoy, A. & Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* **21**, 2850-5.
30. Valdar, W. S. & Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol* **313**, 399-416.
31. Carugo, O. & Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Sci* **6**, 2261-3.
32. Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**, 47-57.
33. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**, 358-61.
34. Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R. & Henrick, K. (2005). E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* **33**, D262-5.
35. Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J. M., John, M., Keller, P. A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Pajon, A., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, G. J., Tagari, M., Tromm, S., Vranken, W. & Henrick, K. (2004). E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* **32**, D211-6.
36. Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res* **28**, 257-9.
37. Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A.,

Ramaswamy, S. & Jap, B. K. (1998). Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science* **281**, 64-71.
38. Gao, X., Wen, X., Yu, C., Esser, L., Tsao, S., Quinn, B., Zhang, L., Yu, L. & Xia, D. (2002). The crystal structure of mitochondrial cytochrome bc1 in complex with famoxadone: the role of aromatic-aromatic interaction in inhibition. *Biochemistry* **41**, 11692-702.
39. Harrenga, A. & Michel, H. (1999). The cytochrome c oxidase from Paracoccus denitrificans does not change the metal center ligation upon reduction. *J Biol Chem* **274**, 33296-9.
40. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *J Mol Biol* **332**, 989-98.
41. Littler, S. J. & Hubbard, S. J. (2005). Conservation of orientation and sequence in protein domain--domain interactions. *J Mol Biol* **345**, 1265-79.
42. Krojer, T., Garrido-Franco, M., Huber, R., Ehrmann, M. & Clausen, T. (2002). Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature* **416**, 455-9.
43. Zheng, J., Knighton, D. R., Xuong, N. H., Taylor, S. S., Sowadski, J. M. & Ten Eyck, L. F. (1993). Crystal structures of the myristylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations. *Protein Sci* **2**, 1559-73.
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-10.
45. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-108.
46. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41.
47. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105**, 1-12.
48. Bolser, D., Dafas, P., Harrington, R., Park, J. & Schroeder, M. (2003). Visualisation and graph-theoretic analysis of a large-scale protein structural interactome. *BMC Bioinformatics* **4**, 45.
49. Park, J., Lappe, M. & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol* **307**, 929-38.
50. Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14**, 309-23.

# Figure Legends

**Figure 1**

(a) Generation of an Assembly from an Entry by application of biologically relevant symmetry operations. The chains are not interacting with one another in the Entry but when the symmetry operation is applied the chains can be seen to interact with other domains. (b) Generation of Assemblies from and Entry by removal of crystal packing artefacts. PQS deems the interaction between the two chains seen in the Entry to be due to a crystal packing artefact and so separates the 2 chains into two Assemblies. (c) Generation of an Assembly or Assemblies from an Entry where both removal of crystal packing artefacts and application of relevant symmetry operations have been applied.

**Figure 2**

The classification of interactions into either PDB interactions or PQS interactions. PQS interactions are interactions where at least one of the chains is a PQS chain. PDB interactions are interactions where both of the chains are PDB chains.

**Figure 3**

The additional non-redundant hetero-fam-pair interactions generated by PQS observed in the Bc1 complex. All the figures are coloured using the same key. All regions which are not directly involved in the interaction are coloured in purple. The two domains of interest which are present in the Entry but are not seen interacting are coloured green and orange. The additional interaction observed in the Assembly is shown between the green and yellow domains. The yellow domain is a copy of the orange domain which has been generated by the application of symmetry operations. The residues which are involved in the interaction

at the interaction surface are coloured in red if they belong to the green domain and blue if they belong to the yellow domain. Figure 3(a) shows the interaction between Cytochrome bc1 domain(a.3.1.3) and Rieske iron-sulfur protein (b.33.1.1) from Entry 1be3[37]. The interaction is of 13 residues in size. Figure 3(b) shows the interaction between Ubiquinol-cytochrome c reductase (f.21.1.2) and Rieske iron-sulfur protein (b.33.1.1) in Entry 1l0n[38]. The interaction is of 12 residues in size.

**Figure 4**

Illustration of the diversity of interaction orientation for interactions which are classified to the same pairwise SCOP family. In this example, the interactions are homo-fam-pair interactions between two prokaryotic protease (b.47.1.1) domains, one of which is a PDB interaction and 6 of which are PQS interactions. For each of the interactions one of the domains in the pair is coloured yellow. All of the domains coloured yellow have been structurally aligned by STAMP and then transposed on to one another. The interacting partners of different colours have been transformed using the same transform as their partner. Since the interactions occur at different orientations, the different coloured domains can be seen to be contacting the yellow domains with different surfaces.

# Table Legends

**Table 1**

The number of additional interactions which are observed by using the Assemblies generated by PQS rather than the Entry for the family level of similarity. The domain interactions are split into those that are seen in the Entry (PDB interactions) and those that are generated by PQS applying symmetry operations (PQS interactions). The row labelled "All Types of Interactions" shows the number of interactions including both PQS and PDB interactions. The row labelled "Additional % Generated by PQS" shows the percentage of additional interactions that are generated by using PQS.

Column 1 shows the number of redundant domain interactions. Then columns labelled "Non-redundant – family level similarity" show the number of additional non-redundant interacting domain pairs generated by PQS using an equivalence based upon family level similarity. Then columns labelled "Non-redundant – orientation within family level similarity" show the number of additional non-redundant interacting domain pairs generated by PQS using an equivalence based family level similarity and also orientation using the iRMSD method of interaction similarity. A "homo-fam-pair" interaction is an interaction between two domains from the same family. Interactions between two domains from different families are classified "hetero-fam-pair" interactions.

**Table 2**

The number of additional interactions which are observed by using the Assemblies generated by PQS rather than the Entry for the superfamily level and 90% sequence identity levels of similarity.

**Figure 1a**

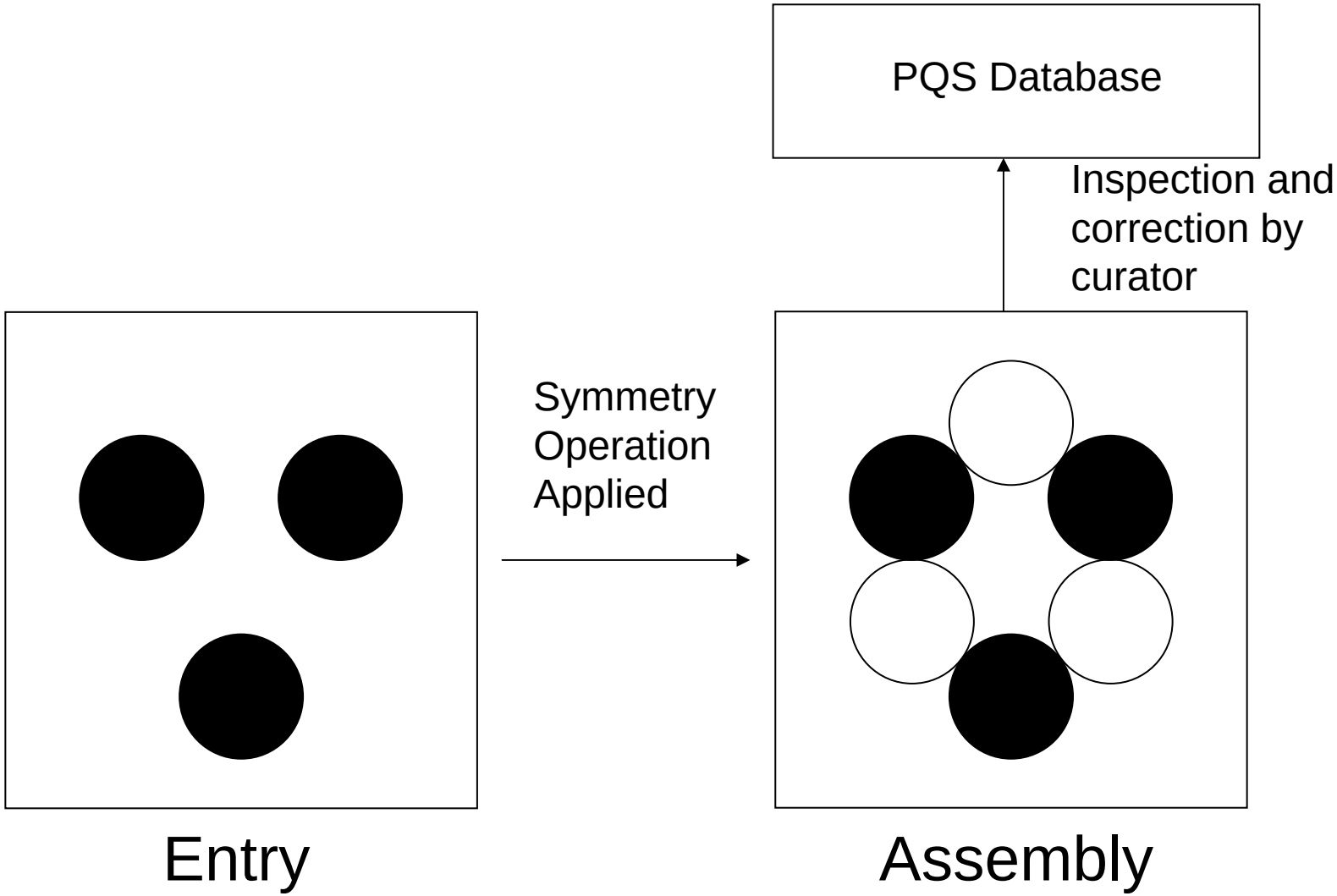
PQS Database
Inspection and correction by curator
Symmetry Operation Applied
Entry
Assembly

**Figure 1b**

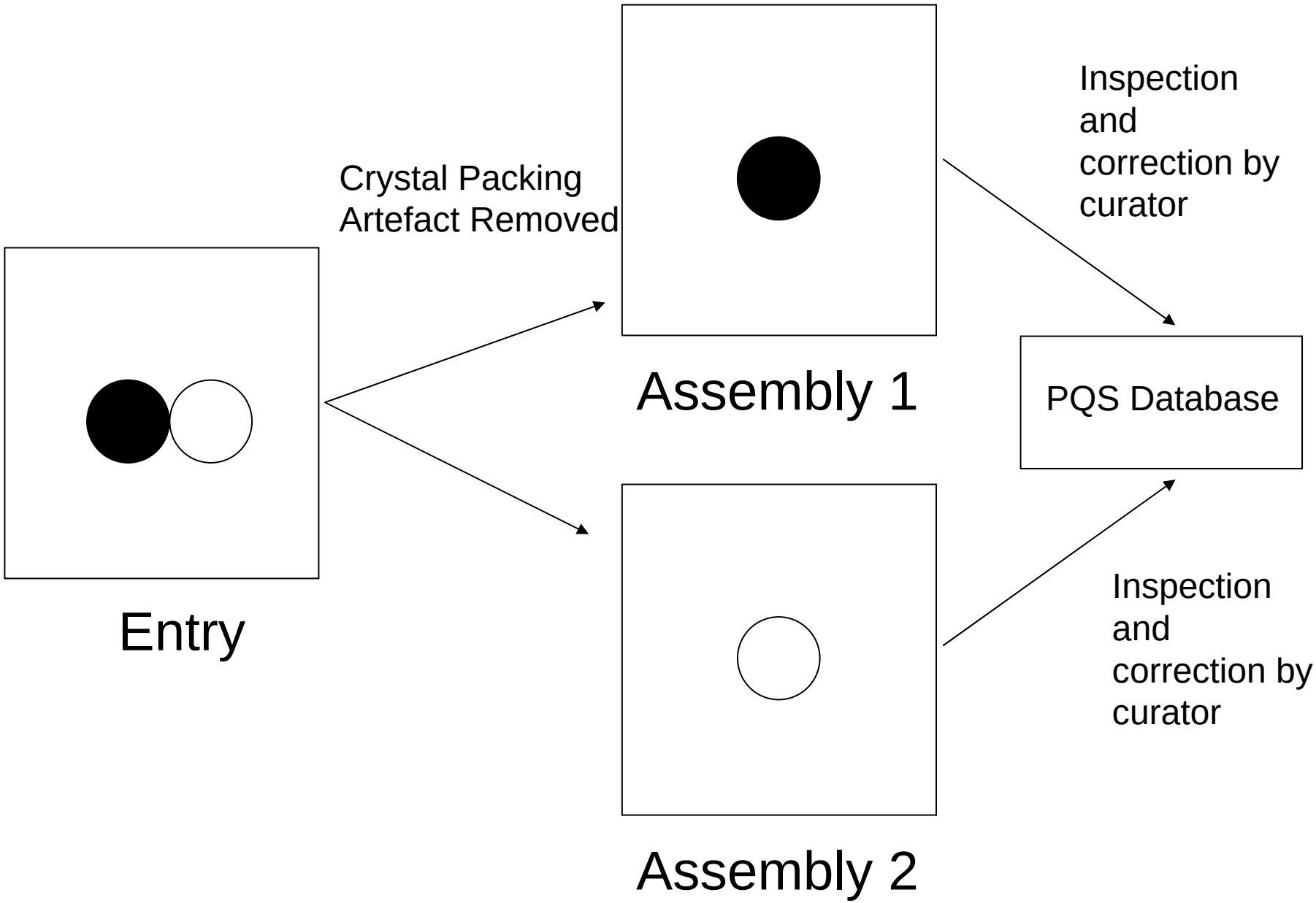
Crystal Packing
Artefact Removed
Inspection
and
correction by
curator
Assembly 1
PQS Database
Entry
Inspection
and
correction by
curator
Assembly 2

**Figure 1c**

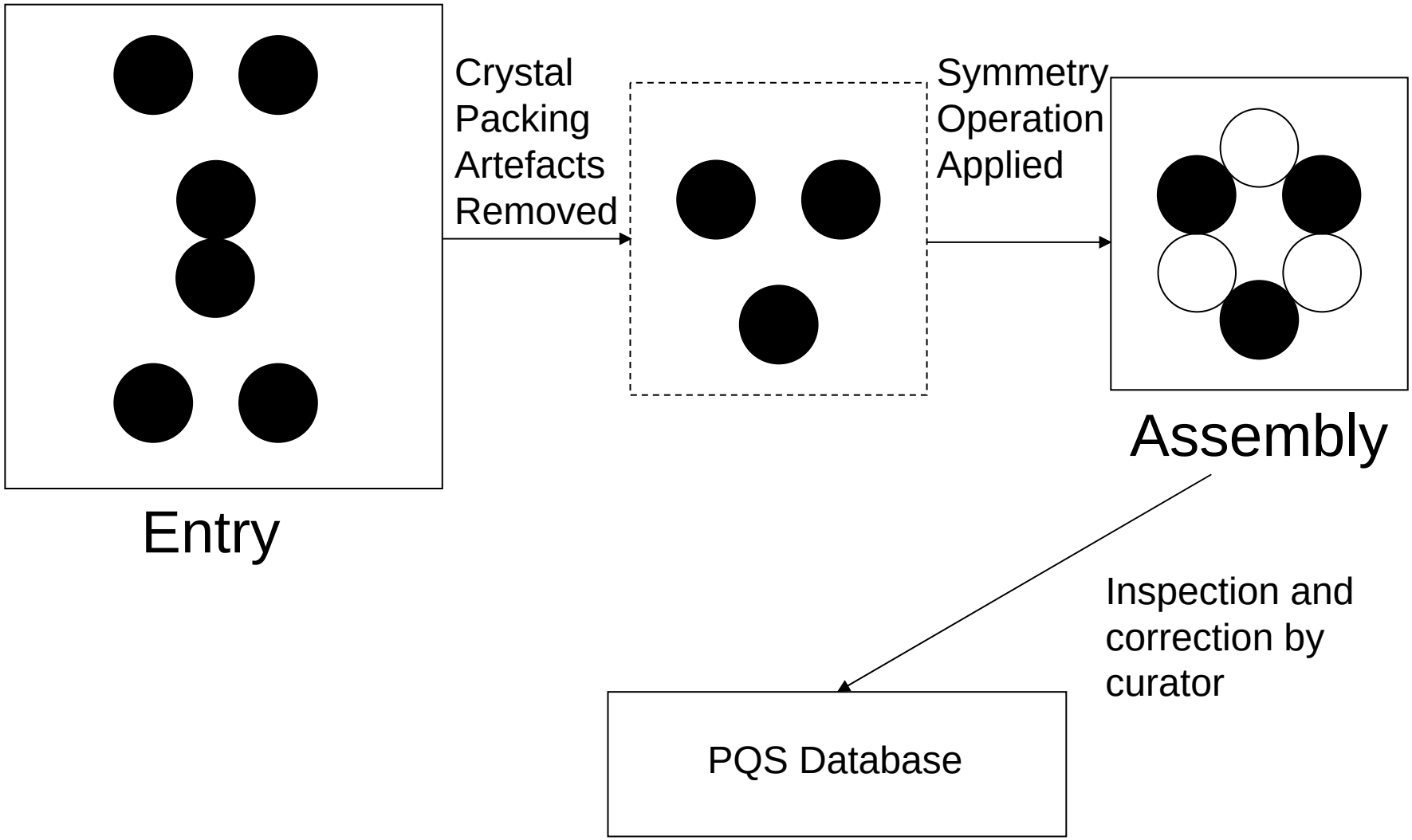
Crystal
Packing
Artefacts
Removed
Symmetry
Operation
Applied
Assembly
Entry
Inspection and
correction by
curator
PQS Database

**Figure 2**

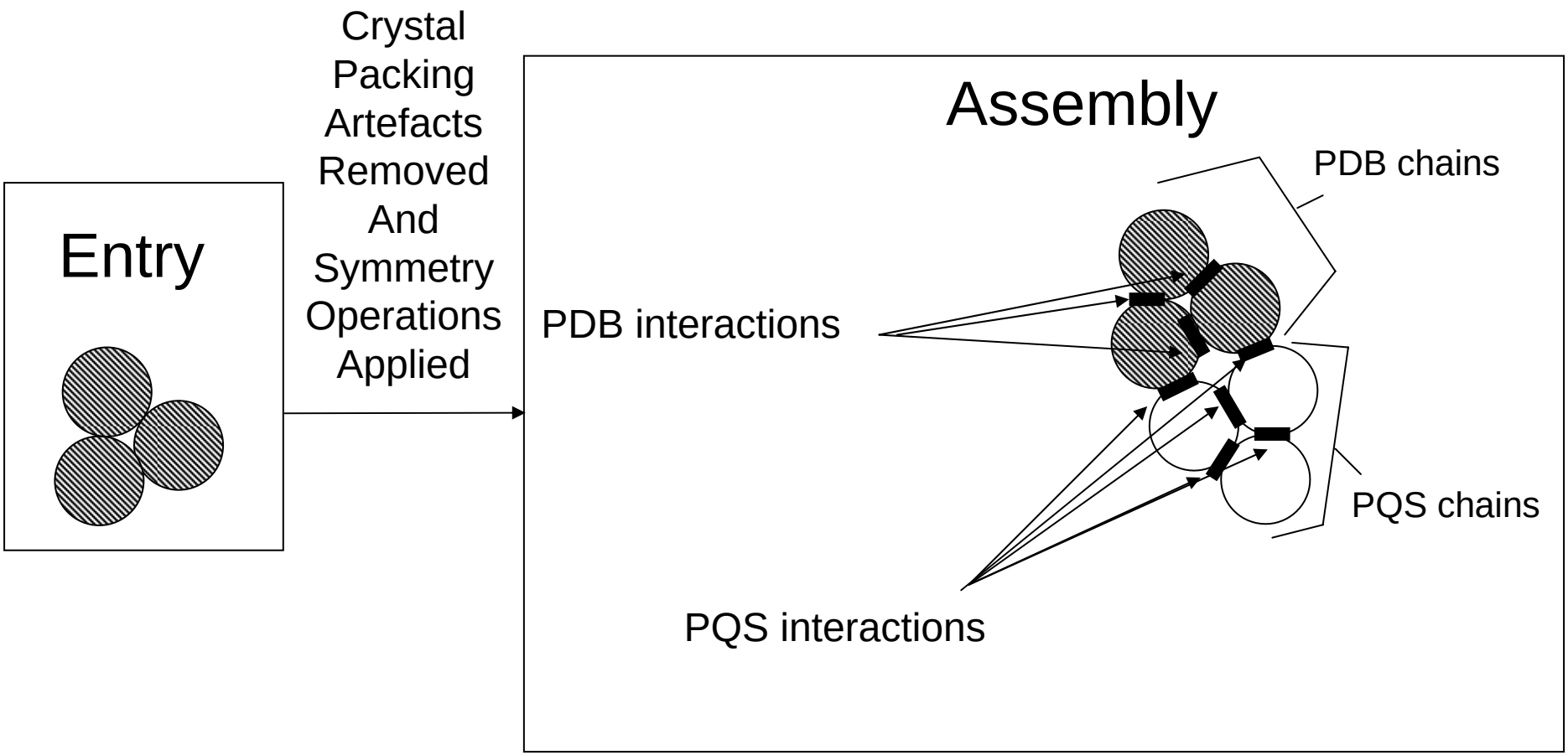
Entry
Crystal Packing Artefacts Removed And Symmetry Operations Applied
Assembly
PDB chains
PDB interactions
PQS chains
PQS interactions

Figure 3a

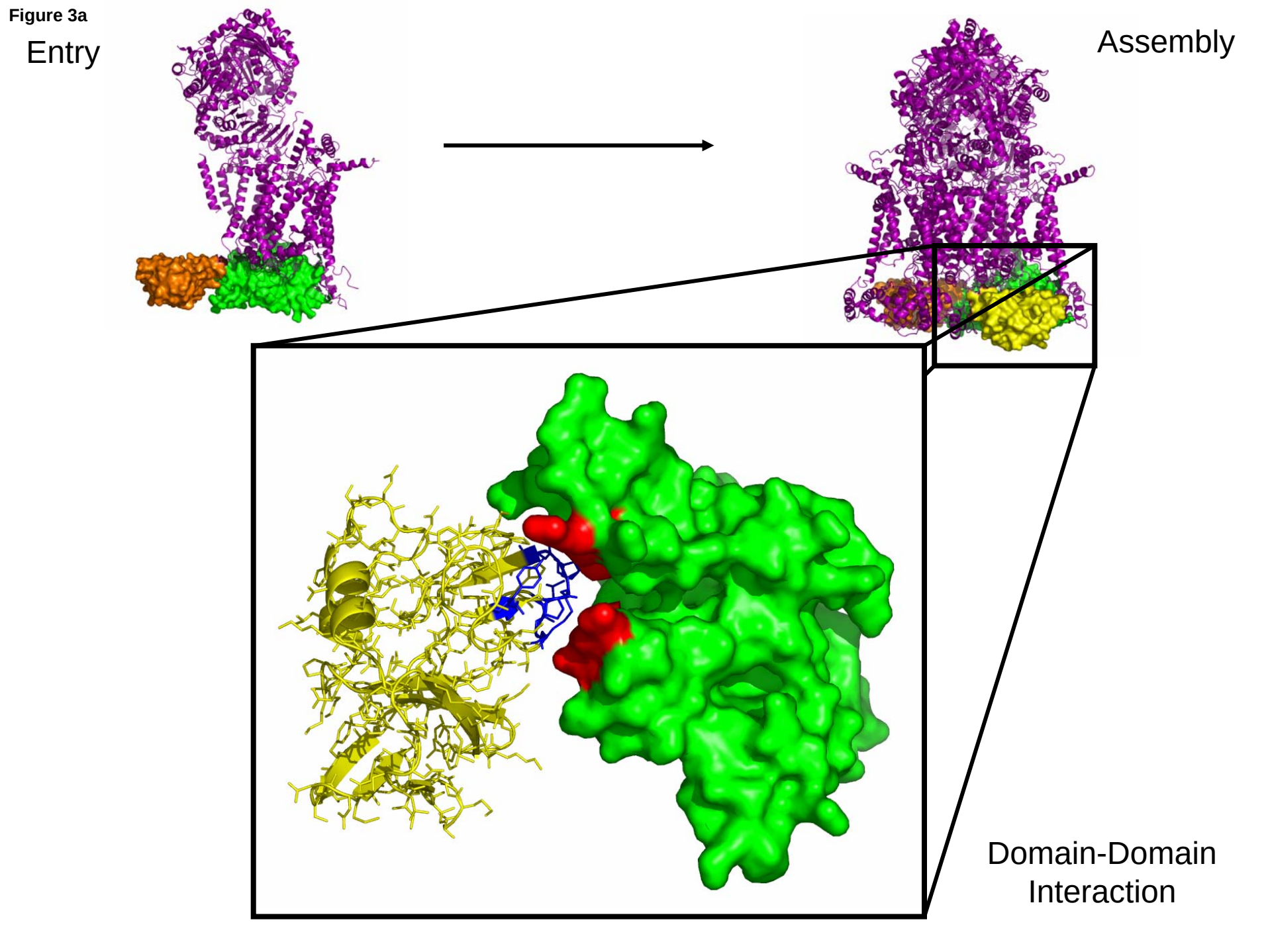

Figure 3b

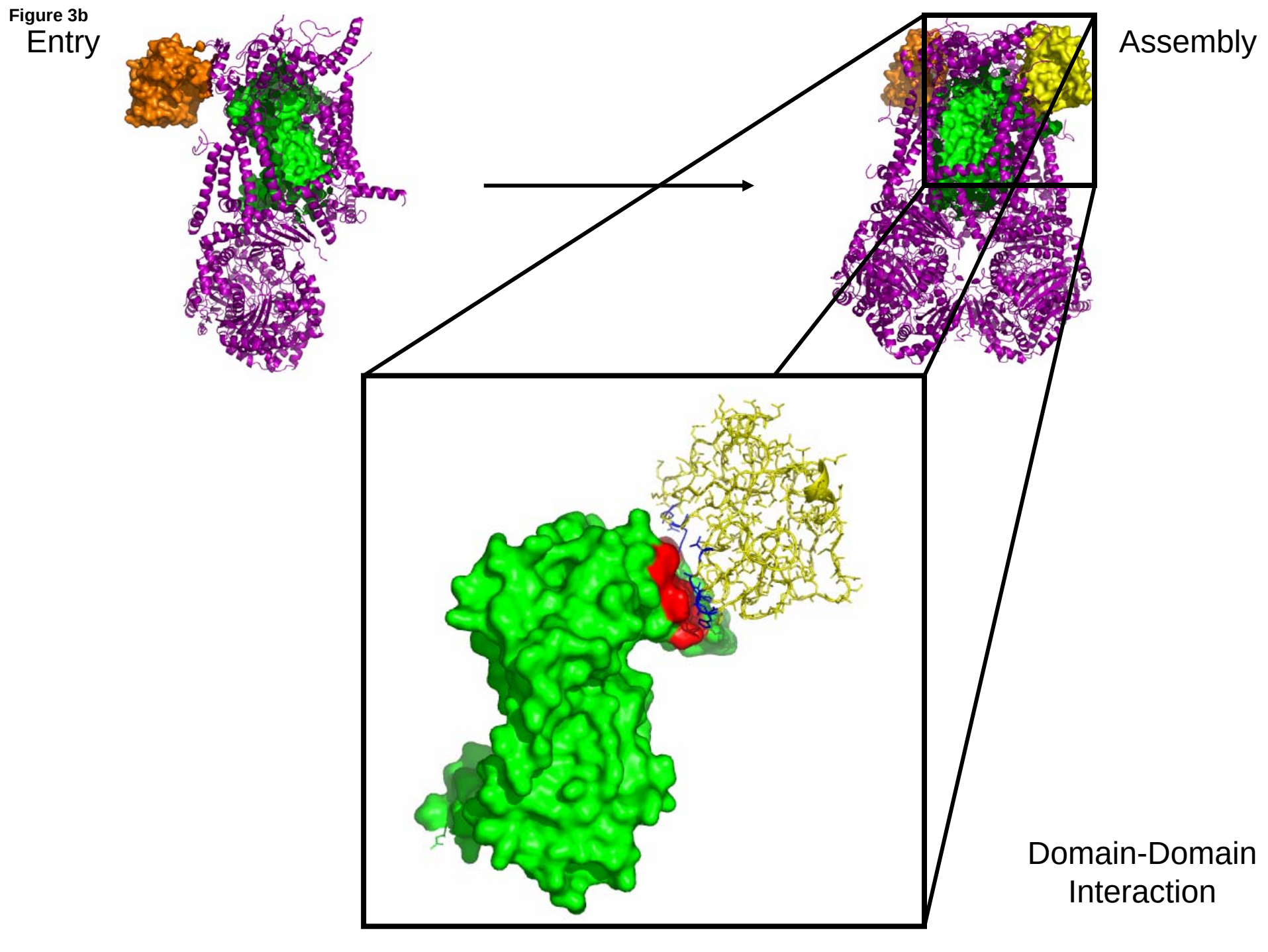

**Figure 4**

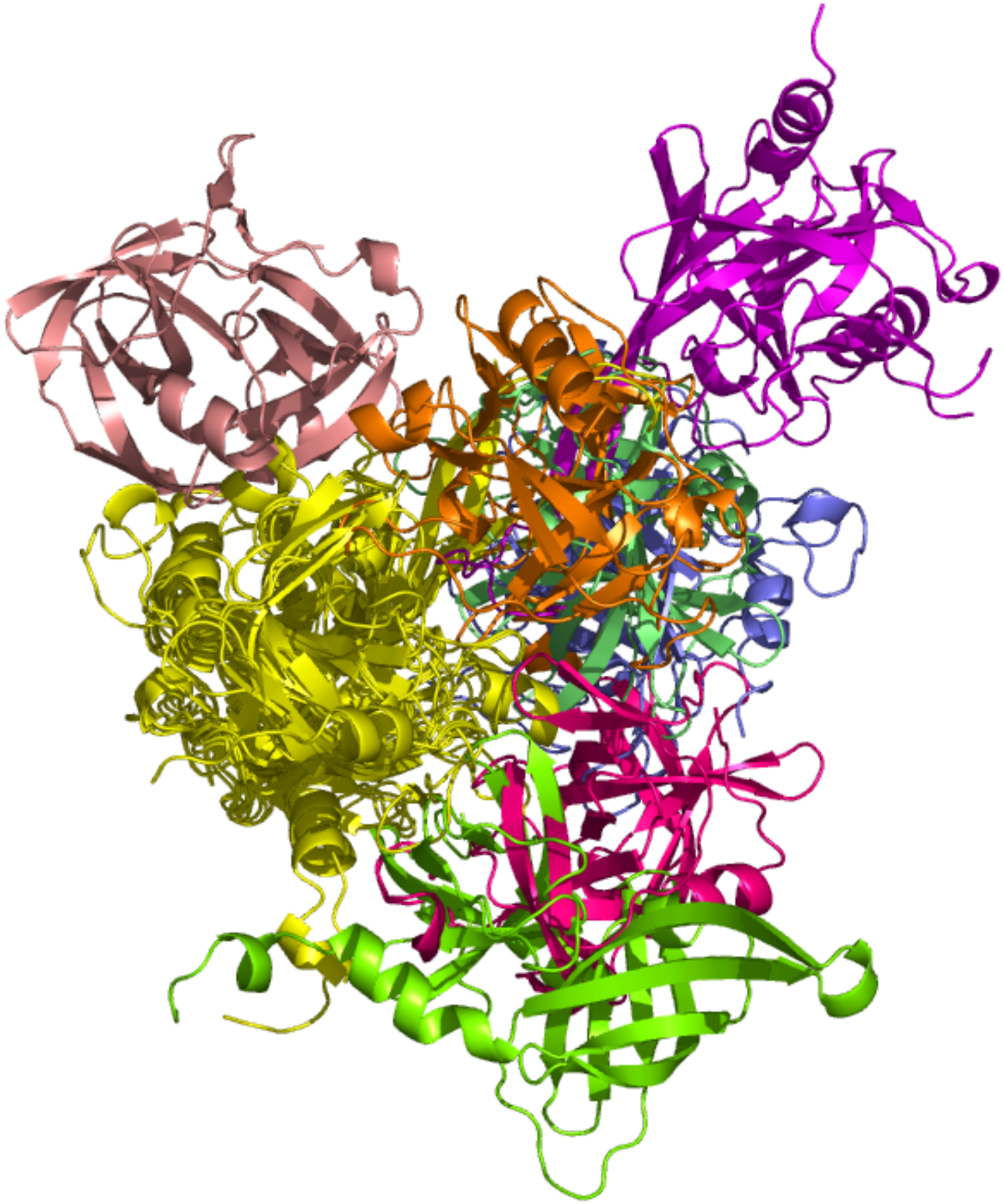

**Table 1**

| | Redundant | Non-redundant – family level similarity | | | Non-redundant – orientation within family level similarity | | |
|---|---|---|---|---|---|---|---|
| | | All pairs | Hetero-fam-pairs | Homo-fam-pairs | All pairs | Hetero-fam-pairs | Homo-fam-pairs |
| All Types of Interactions | 61783 | 2579 | 1177 | 1402 | 5677 | 2025 | 3652 |
| PQS Interactions | 25965 | 302 | 16 | 286 | 1455 | 162 | 1293 |
| PDB Interactions | 35818 | 2277 | 1161 | 1116 | 4222 | 1863 | 2359 |
| Additional % Generated by PQS | 72.5% | 13.3% | 1.4% | 25.6% | 34.5% | 8.7% | 54.8% |

**Table 2**

| | Non-redundant – orientation within superfamily level similarity | | | Non-redundant – superfamily level similarity | | | Non-redundant – sequence similarity | | | Non-redundant – orientation within sequence level similarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All pairs | Hetero-superfam–pairs | Homo-superfam-pairs | All pairs | Hetero-superfam-pairs | Homo-superfam-pairs | All pairs | Hetero-dimers | Homo-dimers | All pairs | Hetero-dimers | Homo-dimers |
| All Types of Interactions | 5468 | 1914 | 3554 | 1894 | 1034 | 860 | 7246 | 3994 | 3252 | 9635 | 4998 | 4637 |
| PQS Interactions | 1369 | 156 | 1213 | 150 | 15 | 135 | 1150 | 71 | 1079 | 2183 | 341 | 1842 |
| PDB Interactions | 4099 | 1758 | 2341 | 1744 | 1019 | 725 | 6096 | 3923 | 2173 | 7452 | 4657 | 2795 |
| Additional % Generated by PQS | 33.4% | 8.9% | 51.8% | 8.6% | 1.5.% | 18.6% | 18.9% | 1.8% | 49.7% | 29.3% | 7.3% | 65.9% |