BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 23 2006, pages 2876–2882
doi:10.1093/bioinformatics/btl490

Structural bioinformatics

ProtBuD: a database of biological unit structures of protein families and superfamilies

Qifang Xu[1,2], Adrian Canutescu[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 and [2]Center for Information Science and Technology, Temple University, 333 Wachman Hall, 1805 N. Broad Street, Philadelphia PA 19122 USA

Received on May 24, 2006; revised on September 12, 2006; accepted on September 20, 2006
Advance Access publication October 2, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Modeling of protein interactions is often possible from known structures of related complexes. It is often time-consuming to find the most appropriate template. Hypothesized biological units (BUs) often differ from the asymmetric units and it is usually preferable to model from the BUs.

Results: ProtBuD is a database of BUs for all structures in the Protein Data Bank (PDB). We use both the PDBs BUs and those from the Protein Quaternary Server. ProtBuD is searchable by PDB entry, the Structural Classification of Proteins (SCOP) designation or pairs of SCOP designations. The database provides the asymmetric and BU contents of related proteins in the PDB as identified in SCOP and Position-Specific Iterated BLAST (PSI–BLAST). The asymmetric unit is different from PDB and/or Protein Quaternary Server (PQS) BUs for 52% of X-ray structures, and the PDB and PQS BUs disagree on 18% of entries.

Availability: The database is provided as a standalone program and a web server from http://dunbrack.fccc.edu/ProtBuD.php.

Contact: Roland.Dunbrack@fccc.edu

1 INTRODUCTION

Current high-throughput methods in proteomics have resulted in substantial information on protein–protein and protein–DNA interactions as well as the contents of large protein complexes (Ito *et al*., 2001). The structures of these interactions are of inherent interest in understanding mechanisms in various pathways and the effects of mutations observed in human populations. When an experimental structure of a complex is not available, computational methods may be used to predict the structure either through *ab initio* docking of the two protein partners or models thereof (Gray *et al*., 2003), or by using known structures of complexes of proteins homologous to the target interacting partners.

It is therefore, valuable to mine as much data on protein interactions as possible from experimental structures and to use these structures as templates for modeling particular target complexes of interest. In recent years, the number of structures in the Protein Data Bank (PDB) (Berman *et al*., 2000) has grown rapidly and the structures have increased in complexity and diversity. The structures of protein complexes, including both homooligomers and heterooligomers, therefore provide a valuable resource for structure prediction and modeling of protein interactions with other proteins, nucleic acids and small molecules.

In order to model particular protein interactions with existing structures, we need to find a template PDB entry or entries with the correct content. That is, given a target sequence (or sequences) we need to find all PDB entries with a protein (or proteins) homologous to the target(s) and then identify those that have the desired interactions. These interactions may include either the correct homooligomer or interactions with proteins of other superfamilies with nucleic acids or small molecules.

Currently, the PDB run by the Research Collaboratory for Structural Bioinformatics (RCSB) provides coordinates for asymmetric units by default in its mmCIF and XML formats. The legacy PDB format files may or may not contain the exact asymmetric unit. The asymmetric unit is defined as the smallest portion of a crystal structure that can be used to build the unit cell (and hence the crystal) using crystal symmetry operators. This is in contrast to the biological unit (BU), which may be larger or smaller than the asymmetric unit and represents a hypothetical biologically active structure. These BUs are built by using symmetry operations from the crystallographic space group to build biological assemblies larger than the asymmetric unit or may be subsets of the asymmetric unit, in which case an asymmetric unit may be broken up into two or more BU files. In some cases, the biological and asymmetric units are identical.

There are currently two main sources for BU information. RCSB provides rules in terms of symmetry operations for building BUs in its mmCIF and XML formatted files, as well as cartesian coordinates in the legacy PDB format. Since January 1, 1999, these BUs are those approved by the authors (Z. Feng *et al*., personal communication). The Protein Quaternary Server (PQS) (Henrick and Thornton, 1998) also provides information on BUs based on analysis of interfaces within crystals and this information sometimes differs from that in the PDB. In PQS (Henrick and Thornton, 1998), the procedure for generating a biological molecule is divided into two steps: selecting protein contacts and filtering out crystal-packing. A potential quaternary assembly is built up recursively by adding monomeric chains based on the number of inter-chain

*To whom correspondence should be addressed.

atomic contacts. Change in surface area and other parameters are used to discriminate between crystal-packing and likely functional protein–protein interactions. Although, BUs of both PDB and PQS are often hypothetical and not supported by direct physical experiments outside of the crystal structure, these data are potential sources of useful information for modeling the BUs of proteins of unknown structures.

Currently several databases provide information on which structures possess interactions between members of two protein superfamilies or families as defined by the SCOP database (Murzin *et al*., 1995). For instance, PIBASE (Davis and Sali, 2005) provides a list of structures for a query of two SCOP superfamily or family designations and provides access to coordinates for each pairwise interaction. Interactions in PIBASE are derived from two sources—the author-approved files provided by PDB in legacy PDB format (e.g. pdb1ylv.ent), which as stated above may or may not be the correct asymmetric unit and PQS files of hypothetical BUs as proposed by the authors of PQS, provided also in legacy PDB format files (e.g. 1ylv.mmol). The emphasis is on characterizing pairwise interfaces in terms of surface area and polar/nonpolar content. Links are provided to visualize the interface. PSIMAP/PSIBASE (Gong *et al*., 2005) also allows for binary searches for two SCOP-defined domains and finds all structures containing interactions between the query domains.

In this paper, we present a database called ProtBuD of the contents of BUs across protein families and superfamilies. Our interest differs from databases, such as PIBASE and is based primarily on locating template structures for homology modeling with specific contents at the level of the BU of structure. This may include particular oligomers of a template as well as interactions with nucleic acids and other ligands. In particular, we show that the proposed BUs in the PDB and/or PQS are different from the asymmetric more than half of the time. Many users assume the standard PDB file that they download is 'the structure' without considering the BU files from these other sources.

Our database shows the asymmetric and BU contents side-by-side, and provides quick download access to the relevant files. Importantly, ProtBuD also provides a comparison of the PDB and PQS BUs for every entry so that users can readily identify whether the PDB and PQS BUs are the same or different both in terms of number of monomers and their orientations and interactions. To our knowledge, this is not available in other databases. We show that PDB and PQS have different BUs ∼20% of the time in terms of numbers of protein monomers and an additional 1% of the time in terms of orientations of monomers within the BUs. It is clear that it may be useful to consult both sources for such information when choosing templates for modeling.

A user can search for a particular SCOP designation or a particular entry or chain in an entry and obtain the asymmetric units, and PDB and PQS BUs of nearly all related proteins in the PDB, as defined by SCOP or PSI–BLAST–reachable relationships. The database also provides information on ligands and nucleic acids for each entry in a query result. Thus, a search on the database provides a simple way of surveying the contents of structures on a family- or superfamily-wide basis for a variety of attributes and the results can be sorted by each of these attributes. Any individual PDB entry can be located in the database, regardless of whether its related structures have been identified by SCOP or PSI–BLAST. The database is very fast, and will be a basis for further development and inclusion within our molecular modeling platform, MolIDE (Canutescu and Dunbrack, 2005). ProtBuD is provided as a standalone program and as a web server, both of which provide user-friendly interfaces. The standalone program has greater functionality.

2 METHODS

2.1 Processing of data files

The data in ProtBuD come from four sources: protein structure files from the PDB in XML format (Berman *et al*., 2000; Westbrook *et al*., 2005), BU coordinate files from PQS (Henrick and Thornton, 1998) in the legacy PDB format, domain classification files from SCOP (Murzin *et al*., 1995) and PSI–BLAST hit files from a non-redundant (100%) PDB database of our lab (Wang and Dunbrack, 2003, 2005). We use the XML *entity_id* and *asym_id* identifiers for all molecules, while the other sources use the author chain ids. The XML files provide a correspondence between these identifiers, although this occasionally presents some ambiguities that can usually be resolved as described below.

PDBML (Westbrook *et al*., 2005) is a part of the uniformity project (Bhat *et al*., 2001) of the PDB. The PDB XML data files preserve the logical data model of the PDB Exchange Data Dictionary (Westbrook and Fitzgerald, 2003). Data can be retrieved quickly from XML files and most software development environments provide libraries to read and write XML files. From the XML files, we retrieve the following data:

(1) the *entity_id* and name for each type of molecule in the structure
(2) for each *entity_id*, the asymmetric unit contents in terms of *asym_ids*; there may be several *asym_ids* for a given *entity_id*
(3) the BU contents consisting of symmetry operators applied to *asym_ids*
(4) for protein and nucleic acid polymers, the author chain ids for each *asym_id* molecule to provide links with other databases such as PQS and SCOP that use the author chain ids; the XML files provide the information that the author chain id's may be blank only for polymer *entity_ids*
(5) information on covalent attachments and modified residues, defined in terms of *asym_ids*, residue numbers and atom names
(6) structural determination data, such as experiment type, space group, transformation matrices for converting to unit cell coordinates, missing residues, resolution and R-factors.

Ligands are not always assigned properly to specific BUs by the PDB. Often when an asymmetric unit is broken up into more than one BU, all of the non-polymer ligands are assigned to the first unit. This is a limitation of the current state of the PDB and may be resolved in future releases of the PDB (J. Westbrook and H. Berman, personal communication).

To compare the BUs provided by the PDB and PQS, we use the legacy PDB format *.mmol files provided by PQS, parsing the 'REMARK 300' fields to match PQS chains and PDB author-designated chains. We use the XML files to provide a translation of the author-designated chain ids used by PQS into *asym_id*s and *entity_id*s. Domain definitions are parsed from the latest version of SCOP classification files: http://scop.mrc-lmb.cam.ac.uk/scop/parse/dir.des.scop.txt_1.69 and its description file, http://scop.mrc-lmb.cam.ac.uk/scop/parse/dir.cla.scop.txt_1.69, available from the SCOP website (Andreeva *et al*., 2004).

As part of our PISCES server, we create a non-redundant set of sequences of proteins in the PDB. We apply a modified PSI–BLAST (Altschul *et al*., 1997) (G. Wang and R. Dunbrack, unpublished data) to each of the sequences of this non-redundant set to search the non-redundant protein sequence database ('nr') available from NCBI (Wheeler *et al*., 2005), to create a position-specific scoring matrix or profile. We then search the entire (redundant) PDB database with these non-redundant profiles and use hits with *E*-value better than 1.0E-10 within ProtBuD.

2.2 BUs comparison

ProtBuD contains information on whether the PDB and PQS BUs are the same or not for each entry. The BUs may differ in content in terms of the number of protein monomers for instance and/or in orientation of the monomers with respect to one another. For comparison of PDB BUs with PQS, we generate coordinates for PDB BUs from the data in the XML files.

The procedure is as follows.

(1) Determine *entity_id*s and *asym_id*s for PDB BU from XML struct_biol_genCategory records.

(2) Determine *entity_id*s and *asym_id*s for PQS BU from 'REMARK 300' records in ∗.mmol files and author/*asym_id*/*entity_id* correspondence from PDB XML file records.

(3) If PDB and PQS BUs consist of different numbers of any polymer entity and the polymer contents of one structure is not a subset of the other one then BUs are different.

(4) Compute unique interfaces in PDB BUs. Compute unique interfaces in PQS BUs. Compute Q scores (see below) for each pair of interfaces (one from PQS and one from PDB BU).

(5) If the numbers of each polymer entity are the same and the numbers of interfaces in PDB and PQS BUs are same and the unique interfaces in the PDB BU can be matched one for one to unique interfaces in the PQS BU with Q > 0.5 then BUs are the same.

(6) Else if one BU is a subset of the other in terms of numbers of each polymer entity and the interfaces in the smaller BU can be matched one for one to an interface in the larger BU with Q > 0.5 then the smaller BU is a substructure of the larger BU.

Here, an amino acid contact is defined as two either Cβ or Cα atoms (for glycines) with distance not greater than 12 Å. A chain contact is defined as two chains that have at least 10 amino acid contacts. An interface is defined as the list of contacts between two contacting chains.

The similarity measurement of interfaces is based on a distance-weighted score. The weights are defined as

$$w_{ij} = \begin{cases} \left(1-\left(\frac{d_{ij}}{D_{off}}\right)^2\right)^2 & if\ d_{ij} < D_{off} \\ 0 & if\ d_{ij} \geq D_{off} \end{cases}, \quad (1)$$

where $D_{off} = 12$ Å and d_{ij} is the distance between two atoms in one structure (one from each of two proteins). If the distance is >12 Å, the weight is 0.

The Q function is then defined as

$$Q = \frac{\sum_{i,j} w_{ij}(-k|e_{ij} - f_{ij}|)}{\sum_{i,j} w_{ij}}, \quad (2)$$

where e_{ij} and f_{ij} are the distances for a particular atom pair in the PDB and PQS BU structure, respectively. We use the value of k for 0.5, derived empirically from a range of values. A Q score is = 1 (within round-off error) if two interfaces are identical.

We have to define which pairs are considered in the sum in Equation 2 and which structure is used to calculate the weight w_{ij}. The procedure for computing the Q score for two interfaces in different BUs is as follows:

(1) Calculate e_{ij} for all contacts in interface A in PDB BU

(2) Calculate f_{ij}, the distances for all contacts in interface B in PQS BU

(3) For each contact *ij* in interface A and interface B, compute w_{ij} from Equation 1 using d_{ij} = minimum (e_{ij}, f_{ij})

(4) Remove contact *ij* from the list of interfaces B (if it is listed there)

(5) For those contacts in interface A, but not in interface B, compute w_{ij} using $d_{ij} = e_{ij}$

(6) For each remaining contact in interface B, compute w_{ij} using $d_{ij} = f_{ij}$

(7) Compute Q score from Equation 2.

2.3 Interfaces and contacts

To aid in the comparison of interfaces for the same PDB entry and across PDB entries, we provide surface area and residue contact information on the interfaces in each BU. Interfaces and residue contacts are computed and stored in a database and due to the size of this database, it is stored on our web server. This database is accessed by either the standalone ProtBuD program or the web server version via a web service.

We use unique interfaces to compare PDB and PQS BUs. Interfaces are distinct if they are composed of proteins with different pairs of *asym_id*s, that is coming from different chains in the asymmetric unit. Interfaces are identical if they consist of chains with the same *asym_id*s and use the same symmetry operators. For interfaces with the same *asym_id*s, but different symmetry operators, the interfaces are distinct if their Q score is <0.95 and identical if Q ≥0.95. The latter situation may arise for instance if PDB and PQS use different symmetry operators on the same two chains of the asymmetric unit, but the resulting structure is identical except for rotation and translation.

Residue contacts for this purpose are defined as any inter-atomic distance between two proteins <6 Å. The surface area of each unique interface is calculated with the program NACCESS (Hubbard *et al*., 1993). 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. A PDB formatted file for each unique interface is also provided within ProtBuD via the web service.

2.4 Implementation

The ProtBuD database functionality was implemented using FireBird relational database server (http://firebird.sourceforge.net/). The database structure was designed to be modular, to avoid unnecessary redundancy and to allow fast queries. The database schema (available on our website) conforms to the Third Normal Form (3NF) under a set of functional dependencies designed to avoid unnecessary data duplication. Functional dependencies are considered standard practice in establishing good database designs (Silberschatz *et al*., 2002). The communication between the application and the database server is performed using the ODBC protocol. The data tables are created dynamically just before data insertion.

In order to optimize the query speed, indices are added to the tables. The best tradeoff between speed and the required disk space was achieved by using composite indices, which take advantage of the leftmost prefixing rule. For instance, a composite index (class, fold, superfamily and family) is added to speed up SCOP code queries. Our database can be divided into five independent modules: SCOP, PDB, PQS, PSI–BLAST Hits and BUs comparison. Each module can be created or updated individually. The whole database is connected by SCOP SunID, PDB entry id, *asym_id* and/or author chain id.

The program that creates, updates and queries the database is written in C# .Net. C# is a programming language that has many similarities with C++ and Java. The ProtBuD database project has two parts: a core library that implements all processing functions and the user interface. The core library is also shared with the web server version of the program. The standalone program has a user-friendly interface and a simple installation procedure. The database can be updated weekly from our website. The embedded FireBird database is completely hidden from the user, so that a database server does not need to be installed and maintained separately by the user. The current standalone version can only be installed in Window OS, although future ports of C# to Linux systems may enable future versions for Linux (see http://www.mono-project.com).

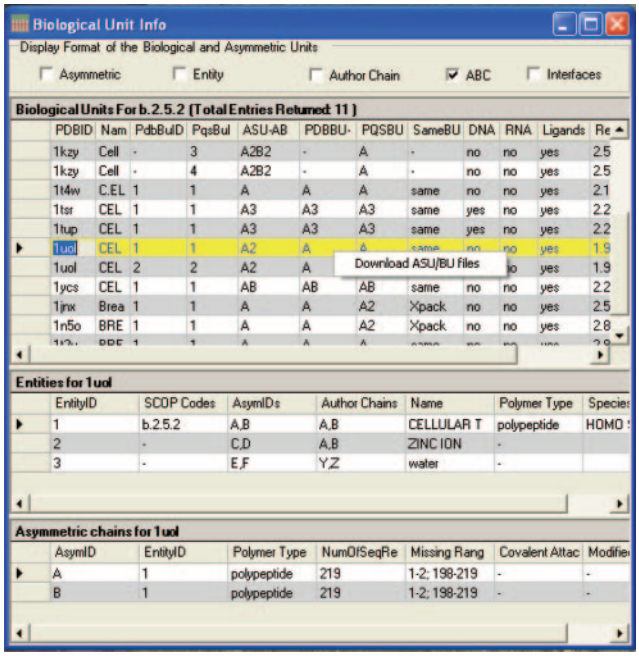

Fig. 1. The asymmetric units and BUs output for SCOP family b.2.5.2 (p53 DNA-binding domain-like family).

3. RESULTS

3.1 Query interface

The central feature of the program tool is the Query. The user enters a PDB entry code with or without a chain identifier and submits the query to the database. The returned SCOP domain definition data are displayed in a data grid. To explore structures with domains in the same family, superfamily or fold, the user clicks the cell with the appropriate SCOP designation. A new window opens and shows the asymmetric units and BUs of all PDB entries with a domain in the same family, superfamily or fold, as shown in Figure 1. Figure 1 gives an example of output from SCOP code input 'b.2.5.2' (p53 DNA-binding domain-like) or PDB entry input '1UOL'.

Four data formats are provided for the asymmetric and BUs: Asymmetric, Entity, Author Chain and ABC formats. The default is the ABC format, which is similar to that used by PQS. The other formats provide more detailed information on which sequences (*entity_id*s) or chains (*asym_id*s) in the asymmetric unit make-up the BUs. In each of these formats, proteins in the asymmetric or BU with the same sequence are placed together in set of parentheses. So, for instance, in the Asymmetric format for a heterotetramer of two different sequences, the form might be (A,B)(C,D), indicating A and B have one sequence and C and D another. If the same structure was an octamer in the BU, the Asymmetric format might be (A2,B2)(C2,D2), indicating that there are two copies of each chain of the asymmetric unit. An alternative octamer might have been (A4)(C4). The difference is important because there may be some structural differences among chains with the same sequence within a single asymmetric unit. The user can show or hide each kind of format by clicking checkboxes at the top of the window. For most purposes, the ABC format is simplest and provides enough information.

To further explore the entities and asymmetric chains of a PDB entry, the user clicks the PDB id cell (leftmost column in Fig. 1), and two tables appear at the bottom of the window. The first of these covers all the entities (by *entity_id*) that are in the asymmetric unit. From this table, the user can get a summary of the kinds of proteins in the asymmetric unit, including their names, SCOP codes and biological species, as well as the identities of other ligands, such as ions and small molecules. The example in Figure 1 shows that 1UOL contains 'CELLULAR TUMOR ANTIGEN P53' and it provides the *asym_id*s and author chain Id's for the proteins in the asymmetric unit. It also indicates that there is zinc and water in the asymmetric unit and the *entity_id* and *asym_id*s used for these. In the lower grid, data are provided for each polymer *asym_id* in the asymmetric unit. The data include the type, the length, the missing residues and modified residues or covalent attachments to these chains.

The user can browse through the PDB entries in the family, superfamily or fold returned by the query by clicking on an entry in the PDB id column in the top window of Figure 1 and using the up or down arrow keys. Searching for an entry with a specific type of ligand, such as ATP, within the structures in a particular superfamily or family can be accomplished easily by navigating up and down the top table and examining the *entity_id* table that appears below as each PDB entry is selected.

The user can also download the coordinate files from the PDB and PQS ftp servers by right-click on the selected rows or cells. Selecting multiple rows is a shortcut to download ASU/BU files for multiple entries. Selecting a single cell, only downloads the ASU or BU file for that cell. The compressed files are decompressed after being downloaded.

If an input PDB entry is not in SCOP, a list of PSI–BLAST hits is returned with *E*-values, percent identities and residue ranges from the PSI–BLAST alignments. Those in SCOP are listed with their SCOP designations. Any of these hits can be clicked to reveal a new window with the asymmetric and BU data. A right-click will produce a BU table with all of the hits listed.

If the interface checkbox is checked, unique interfaces are listed in a new window. Clicking the interface identifier id will display all similar interfaces and their symmetry operators as well as the contacts. The coordinate file for an interface can be downloaded by right-click on the selected rows and cells.

A query may also consist of two different SCOP codes so that a user may obtain all structures that contain members of two different SCOP families, superfamilies or folds. We have not tested whether the two SCOP domains are in fact in contact with each other. It may be in some cases useful to find structures that contain two SCOP domains, even if they are not in contact. They may be in the same protein chain with a linker long enough to separate them, but such a template may still be useful for modeling. Information on SCOP domain contacts can be obtained from other databases, such as PIBASE and PSIMAP. If a user does not know the SCOP codes, these can be obtained by single queries to our database with PDB entry identifiers and then combining the results in the dual SCOP query.

3.2 Comparison of PDB and PQS BUs

Analysis of the current database provides some useful information on the state of our knowledge of BUs and the contents of the PDB and PQS databases. The current database contains 38 477 PDB

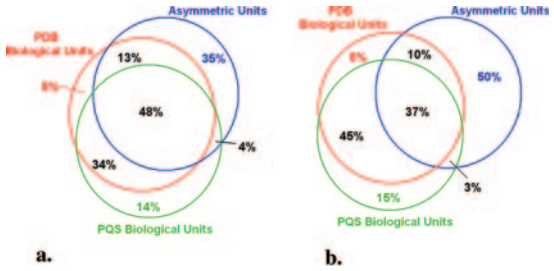

Fig. 2. Venn diagrams of asymmetric units, PDB BUs and PQS BUs. (**a**) Percentages are in terms of 27 996 entries common between PDB and PQS. (**b**) Percentages are in terms of 36 619 BUs. The overlaps are based on the percentage of entries or BUs that have the same contents, in terms of number of each type of polymer chain.

entries, 47 716 PDB BUs, 25 970 SCOP entries, 3114 SCOP families and 40 121 PQS BUs (on 30 275 PDB entries) including 1481 crystal-packing interactions and 22 382 distinct PSI–BLAST query entries. In SCOP, ~55% of PDB entries (14 331) have two or more domains, of which 41% of these entries (5593) have two or more domains in different SCOP families, indicating a great deal of information on protein domain interactions of interest in modeling. These interactions have been analyzed by others (Davis and Sali, 2005; Gong *et al*., 2005; Park *et al*., 2005). Our database also provides information on DNA/RNA and ligands, which are inconsistently annotated within SCOP. For instance, only about one-third of proteins bound with DNA are annotated as such in SCOP. In the PDB, 7% of proteins are complexed with DNA or RNA (2637), while 64% of proteins (24 671) contain other non-covalent ligands (excluding water).

Due to the asymmetric unit and BU representation formats that we use, we can easily compare asymmetric units and BUs across families and superfamilies and between PDB and PQS. We first compared PDB and PQS BUs in terms of their *entity_id* formats, i.e. whether they contain the same number of copies of each protein. Figure 2 shows the differences between asymmetric units and BUs provided by PQS and PDB across the entire archive. We excluded entries that had missing BU information in PDB or PQS (mostly NMR structures, which are not covered in PQS). The resulting 27 996 entries are analyzed. Each circle in the Venn diagram therefore represents the same 27 996 entries. Two Venn diagrams are provided. In Figure 2a, the percentages are in terms of total number of entries, while in Figure 2b, the percentages are in terms of total numbers of BUs.

The asymmetric unit content is different from the PDB or PQS BUs (or both) for 52% of entries (Fig. 2a: 35% different from both PDB and PQS; 13% different from PQS but same as PDB and 4% different from PDB but same as PQS). The asymmetric unit and PDB BUs are different for 53% of all BUs defined by RCSB (Fig. 2b). Of these, the BU is smaller than the asymmetric unit for 38% and larger for 15% of all BUs. The asymmetric unit and PQS BU are different for 60% of all BUs (Fig. 2b) and the PQS BU is smaller than the asymmetric unit for 38%, of entries and larger for 22%.

We compared the BUs from PDB and PQS for both content and orientation of the macromolecules contained in them. The column SameBU in Figure 1 indicates if two BUs are the same or not, as detailed in Table 1. Two BUs are the same if they contain the same polymer entities with the same number and types of interfaces. Two BUs with the same entity contents may be different either because of a different number of interfaces, marked by 'difNum' or because of different interaction orientations between proteins marked by 'difOrient'. An example of difNum is shown in Figure 3a and b (PDB entry 1TUI) for PDB and PQS BUs, respectively, while an example of difOrient is shown in Figure 3c and d (PDB entry 1B6R), again for PDB and PQS, respectively. DifNum entries may also have different orientations as well as different numbers of interfaces. PQS labels some BUs as 'XPACK' and describes them as probably due to crystal-packing but nevertheless of possible interest. There are currently 1220 such XPACK BUs and these are labeled 'XPACK' in the SameBU column.

PDB and PQS agree on BUs for 82% of the entries based on *entity_id* content (Fig. 2a: 48% where they are both the same as the ASU and 34% where they agree with each other but are different from the ASU). We further examined these entries to determine if the BUs contain the same number and kinds of interactions, i.e. whether they were in fact the same structure. For 18 247 BUs with same entity format and more than one chain, the interfaces in one BU were compared to the interfaces of the other using a distance-weighted similarity score as described in methods. Currently in ProtBuD, 325 BUs (1.8%) are different either in the interface number or in the relative orientation of the proteins. We visually checked all different BUs automatically returned from the programs and found seven false negatives due to residue numbering and chain matching problems in PQS, which we have manually corrected in the database. We calculated the percentages shown in Table 1 for only those entries submitted since January 1, 1999 and the results were very similar (data not shown).

We also investigated those BUs where one BU is a subset of the other in terms of the *entity_id* content. For instance, for entry 1A4P, the PDB BU has two copies of entity 1 while the PQS BU is a homotetramer. We analyzed 1821 pairs of BUs where one was an entity subset of the other, excluding entries for which one of the BUs was a monomer. For 94.7% of these BUs, one BU is a substructure of the other.

3.3 Limitations of the data

The PDB and PQS BUs are only hypothetical, since rarely have the proteins been studied by the appropriate physical experiments in solution. Even when the physical size of the BU may be known (as a dimer or tetramer for instance from analytical centrifugation or native gels), the actual physical interfaces are not. A startling example of this is the sulfotransferase family for which Petrotchenko *et al*. experimentally determined the dimer interface using crosslinking, mass spectrometry and mutational analysis (Petrotchenko *et al*., 2001). Many of the sulfotransferases (11 different family members in 21 PDB entries in 12 different space groups) are labeled as monomers in PQS and PDB, and those that are dimers are not the same dimer as identified experimentally with only three exceptions in the PDB BUs. However, visual examination of all crystal contacts for these structures indicates that the Petrotchenko dimer is present in all of them (data not shown). The BUs reported by PDB, PQS and ProtBuD therefore should not be taken as certain but as possibilities for modeling in different oligomeric forms for particular proteins of interest.

To examine this further, we looked at particular sequences that appear in multiple PDB entries as monomers or homooligomers to

Table 1. Flags for SameBUs column in Figure 1

Flags	Descriptions	Example	%[a]
Same	Same entity contents, same orientation	1GZH: PDBBU-Entity: (1.1)(2.1) PQSBU-Entity: (1.1)(2.1) Interfaces in PDBBU and PQSBU are same	82
DifNum	Same entity contents, different number of interfaces	1TUI: PDBBU-Entity: (1.3) PQSBU-Entity: (1.3) 1tui.pdb1 has 2 interfaces and 1tui_1.mmol only 1	0.4
DifOrient	Same entity contents, same number of interfaces but different orientations	1B6R: PDBBU-Entity: (1.2) PQSBU-Entity: (1.2) The number of interfaces in both BUs is 1. However their orientations are different.	0.6
Substruct	Entity content of one BU is a subset of the other one. Interfaces in smaller **BU** are contained in the larger **BU**	1B71: PDBBU-Entity: (1.2) PQSBU-Entity: (1.4) The PDB dimer is contained within the PQS tetramer	14
Dif	Different entity contents and one structure is not a substructure of the other	1A4P: PDBBU-Entity: (1.2) PQSBU-Entity: (1.4) PDB **BU** is not the same size nor a substructure of the PQS BU	0.3
Xpack	PQS XPack	1B88: PDBBU-Entity: (1.2) PQSBU-Entity: (1.2)	3.3

[a]Total number of BUs: 36 619.

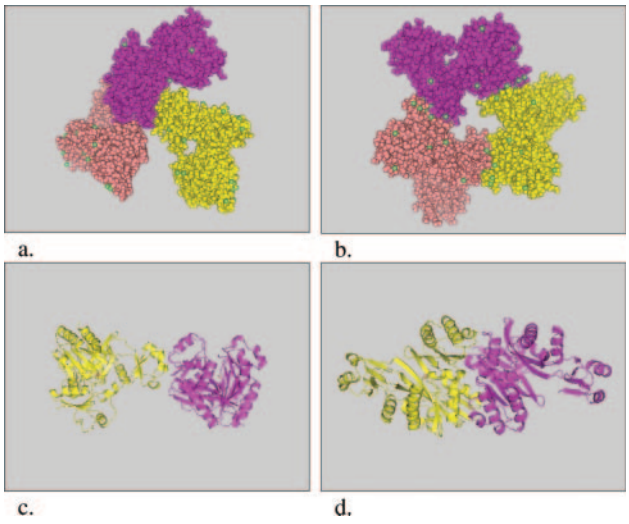

Fig. 3. Two examples of BU differences between PDB and PQS. (**a**) PDB BU for 1TUI. (**b**) PQS BU for 1TUI. (**c**) PDB BU for 1B6R. (**d**) PQS BU for 1B6R. 1TUI is an example of difNum, where the number of interfaces differs between the two BUs. 1B6R is an example of difOrient, where the orientations of the two monomers in the BUs are different.

see if PQS or PDB reported different BUs. We examined all unique polypeptide sequences in the PDB that appear in multiple entries without other polypeptide or polymer sequences. We focused on sets of structures for a particular sequence with the same space group and crystal dimensions (within 5%), so that in fact the crystal forms for the protein are the same. An example for the $P3_221$ crystal form of the P21 protein is shown in Table 2. All of the crystals have the same space group and same dimensions ($a = b = 40.2 \pm 0.6$ Å. $c = 160.4 \pm 2.0$ Å) and angles ($\alpha = \beta = 90$, $\gamma = 120$). Yet, in the PDB, six of the structures are monomers and five are dimers. In PQS, seven of the structures are dimers, although five of these are labeled as 'XPACK'.

We found 3057 such sequences in 9619 different structures with BU information in the PDB. In theory, we would expect the BUs for the same protein in different structures but with the same crystal forms to be identical. However, a total of 193 of these sequences (6%) involving 857 structures (9%) contained different PDB BUs across the entries for each sequence. For PQS, we found 2447 sequences involving 7600 structures with BU information from multiple structures in the same space group. A total of 244 of these sequences (10%) involving 978 entries (13%) exhibited more than one BU form. While it is possible to examine differences in BUs for proteins in different crystal forms, in these cases it is possible that crystallization conditions (pH, temperature and ligands) might change the multimerization state in biologically meaningful ways. Therefore, we did not analyze differences in BUs across different crystal forms. This will be performed in future work.

Table 2. BUs for the $P3_221$ crystals of P21

PDB	PDB BU	PQS BU
121P	A2	A2 (XPACK)
1CTQ	A	A2
1GNP	A2	A2 (XPACK)
1GNQ	A2	A2 (XPACK)
1GNR	A2	A2 (XPACK)
1P2S	A	A
1P2T	A	A
1P2U	A	A
1P2V	A	A
1QRA	A	A2
5P21	A2	A2 (XPACK)

BUs given in ABC format, such that A is a monomer, A2 is a homodimer and A2 (XPACK) is a crystal-packing interface indicated by PQS to be of possible interest. The asymmetric unit is a monomer.

4 DISCUSSION

Protein interactions play a key role in carrying out a cell's biological functions. These interactions include both homo and

heteromultimeric structures. Our database of BUs can be queried with one or more protein sequences to obtain a list of PDB entries with domains from each sequence. This is the first step in enabling modeling biological systems with greater complexity than the modeling of single proteins, available from many protein modeling servers. We intend to make ProtBuD an integral part of our graphical user interface for protein homology modeling, MolIDE (Canutescu and Dunbrack, 2005), so that proteins can be modeled as part of homo or heterooligomeric complexes with the inclusion of important ligands and residue modifications.

The main purpose of our database is to provide information on the content of BUs for identifying potential templates for predicting the structures of protein complexes when combined with MolIDE. One of the main features of ProtBuD is that with a single query a user can find information on BUs and ligands across a family or superfamily. Normally, one performs a PSI–BLAST search of the PDB and then must look up this information for each returned hit manually one by one. This is a tedious process and ProtBuD makes it much easier.

But the data have inherent interest when viewed across families or superfamilies. A major result of the analysis of data in ProtBuD is first of all that PQS and PDB agree on only 82% of the BUs of X-ray structures, indicating that there is considerable uncertainty of what these structures really are. This provides an opportunity for further analysis and modeling of proteins using a number of hypothesized multimeric structures to be used for further experimental testing. Second, even within a single family or superfamily, there are often different BUs available, which may or may not be correct, but again allowing one to use each as a possible template for model building and for possible further testing with carefully designed experiments, Third, our database allows a user to find templates with specific ligands such as RNA or DNA, or ions or small molecules. Knowledge of all these interactions—homomultimer, heteromultimer, nucleic acids and ligands—are all key to understanding experimental data and the biological effects of mutations that may exist in the population.

ACKNOWLEDGEMENTS

The authors wish to thank Brian Weitzner and Rajib Mitra for visual comparison of BUs. Support was provided by the NIH and the Pew Charitable Trusts (to the Molecular Modeling Facility of Fox Chase Cancer Center) and the Pennsylvania Tobacco Settlement. Funding to pay the Open Access publication charges for this article was provided by NIH Grant R01 GM 73784 to R.L.D.

Conflict of Interest: none declared

REFERENCES

Aloy,P. *et al*. (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol*., **332**, 989–998.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI–BLAST: a new generation of database programs. *Nucleic Acids Res*., **25**, 3389–3402.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Bhat,T.N. *et al*. (2001) The PDB data uniformity project. *Nucleic Acids Res*., **29**, 214–218.

Canutescu,A.A. and Dunbrack,R.L., Jr (2005) MollDE: a homology modeling framework you can click with. *Bioinformatics*, **21**, 2914–2916.

Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*.

Gong,S. *et al*. (2005) PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*, **21**, 2541–2543.

Gray,J.J. *et al*. (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol*., **331**, 281–299.

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

Hubbard,S. and Thornton,J. (1993) 'NACCESS', Computer Program. Department of Biochemistry and Molecular Biology, University College, London.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Park,D. *et al*. (2005) Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interactome map). *Bioinformatics*, **21**, 3234–3240.

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

Silberschatz,A., Korth,H.F. and Sudarshan,S. (2002) *Database system concepts*. McGraw-Hill, New York.

Wang,G. and Dunbrack,R.L., Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,G. and Dunbrack,R.L., Jr (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*., **33**, W94–W98.

Westbrook,J. *et al*. (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.

Westbrook,J.D. and Fitzgerald,P.M. (2003) The PDB format, mmCIF, and other data formats. *Methods Biochem. Anal*., **44**, 161–179.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.