BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 21 2006, pages 2612–2618
doi:10.1093/bioinformatics/btl447

Structural bioinformatics

DOCKGROUND resource for studying protein–protein interfaces

Dominique Douguet[1,*], Huei-Chi Chen[2], Andrey Tovchigrechko[3] and Ilya A. Vakser[3,4]

[1]Centre de Biochimie Structurale, CNRS, U5048, Université Montpellier 1, INSERM, U554, 29, rue de Navacelles, Montpellier, F-34090, France, [2]Department of Applied Mathematics and Statistics, Math Tower 2-109, SUNY Stony Brook, Stony Brook, NY 11794-3600, USA, [3]Center for Bioinformatics and [4]Department of Molecular Biosciences, The University of Kansas, 2030 Becker Drive, Lawrence, KS 66047-1620, USA

Received on May 5, 2006; revised on July 29, 2006; accepted on August 16, 2006
Advance Access publication August 23, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Public resources for studying protein interfaces are necessary for better understanding of molecular recognition and developing intermolecular potentials, search procedures and scoring functions for the prediction of protein complexes.

Results: The first release of the DOCKGROUND resource implements a comprehensive database of co-crystallized (bound–bound) protein–protein complexes, providing foundation for the upcoming expansion to unbound (experimental and simulated) protein–protein complexes, modeled protein–protein complexes and systematic sets of docking decoys. The bound–bound part of DOCKGROUND is a relational database of annotated structures based on the Biological Unit file (Biounit) provided by the RCSB as a separated file containing probable biological molecule. DOCKGROUND is automatically updated to reflect the growth of PDB. It contains 67 220 pairwise complexes that rely on 14 913 Biounit entries from 34 778 PDB entries (January 30, 2006). The database includes a dynamic generation of non-redundant datasets of pairwise complexes based either on the structural similarity (SCOP classification) or on user-defined sequence identity. The growing DOCKGROUND resource is designed to become a comprehensive public environment for developing and validating new methodologies for modeling of protein interactions.

Availability: DOCKGROUND is available at http://dockground.bioinformatics.ku.edu. The current first release implements the bound–bound part.

Contact: douguet@cbs.cnrs.fr

1 INTRODUCTION

The cellular machinery is based on the network of intermolecular interactions. The knowledge of structural information on protein–protein interactions is fundamental to understanding protein function. It is also an essential step in correcting biological dysfunction related to diseases. The experimentally solved protein–protein complexes represent only a fraction of protein–protein complexes existing *in vivo*. Thus most of the protein–protein interactions have to be characterized by computational modeling (Russell *et al*., 2004). The computational approaches benefit from the information resulting from multiple sequenced genomes. In the post-genomic era, the software dedicated to structural modeling of protein interactions (Marshall and Vakser, 2005; Vajda, *et al*., 2002) plays an increasingly important role in the emergent field of 'interactome'.

Although the structure of protein–protein complexes is generally more difficult to determine than the structure of individual proteins, the number of experimentally determined complexes is statistically significant. The databases of protein–protein complexes are indispensable for systematic studies of protein interactions and the design of new predictive tools. Our previous dataset of protein–protein complexes was built by Vakser and Sali (unpublished data) based on 1997 release of PDB containing 5013 entries. Since its release it has been extensively used in studies of knowledge-based potentials (Glaser *et al*., 2001), intermolecular energy landscapes (Papoian and Wolynes, 2003; Tovchigrechko and Vakser, 2001; Vakser *et al*., 1999), docking methodology (Tovchigrechko *et al*., 2002) and other studies. Some datasets of protein–protein complexes have been compiled and used to address various aspects of physicochemical and structural features of protein–protein interfaces (Bogan and Thorn, 1998; Dasgupta *et al*., 1997; Keskin *et al*., 1998; Keskin *et al*., 2004; Larsen *et al*., 1998; Lijnzaad and Argos, 1997; Lo Conte *et al*., 1999; Lu *et al*., 2003; Ponstingl *et al*., 2000). Most existing databases are either non-comprehensive or not automatically updated or fully querying. The DOCKGROUND resource is regularly updated, filtered and annotated. Our datasets have options to exclude particular complexes (ligands at the interface, disulphide bonds and alternative binding modes) as well as redundancies based on sequence or structural similarities. At the same time, a user can access the full (redundant) set of structures (e.g. to study structural variability of the interface among homologous complexes). The first DOCKGROUND release implements the database of co-crystallized (bound) protein–protein complexes and provides the foundation for the future expansion to unbound (experimental and simulated) protein–protein complexes, modeled protein–protein complexes and systematic sets of docking decoys. The growing DOCKGROUND resource is designed to become a comprehensive public environment for developing and validating new methodologies for modeling of protein interactions.

2 SOURCE OF QUATERNARY STRUCTURES

The DOCKGROUND dataset was originally built on the basis of the PDB release containing >34 000 entries (January 2006). When crystallographic structures are deposited to PDB, the primary

[*]To whom correspondence should be addressed.

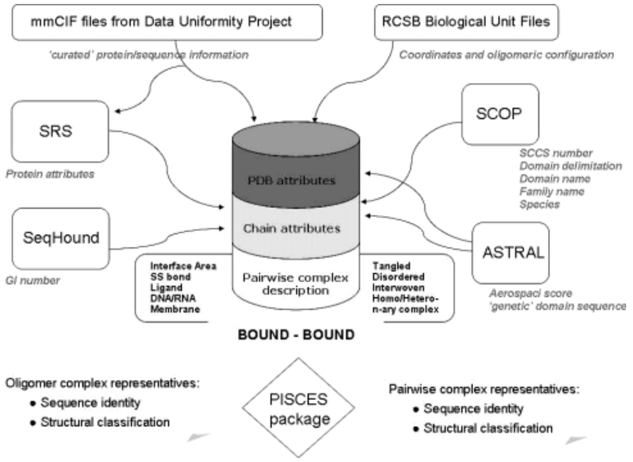

mmCIF files from Data Uniformity Project
'curated' protein/sequence information
RCSB Biological Unit Files
Coordinates and oligomeric configuration
SRS
Protein attributes
SeqHound
GI number
SCOP
SCCS number
Domain delimitation
Domain name
Family name
Species
ASTRAL
Aerospaci score
'genetic' domain sequence
PDB attributes
Chain attributes
Pairwise complex description
Interface Area
SS bond
Ligand
DNA/RNA
Membrane
Tangled
Disordered
Interwoven
Homo/Hetero-n-ary complex
BOUND - BOUND
PISCES package
Oligomer complex representatives:
• Sequence identity
• Structural classification
Pairwise complex representatives:
• Sequence identity
• Structural classification

Fig. 1. Schematic representation used in the DOCKGROUND database construction. Primary data source and external programs are shown in black and white. DOCKGROUND databases are shown in gray.

(original) coordinate file generally contains one asymmetric unit (a.s.u.). An a.s.u. is the smallest portion of the crystal structure to which crystallographic symmetry can be applied to generate one unit cell. The unit cell is the smallest unit in a crystal, which upon translation in three dimensions makes up the entire crystal. The a.s.u. is used by the crystallographer to refine the structure against experimental data and does not necessarily represent a biologically functional molecule. Depending on the a.s.u., the spacegroup symmetry operations consisting of either rotations or translations must be performed to obtain the complete biological unit. Thus a biological unit may be built from one copy of the a.s.u., multiple copies of the a.s.u. or a portion of the a.s.u. (http://www.rcsb.?org/robohelp_f/data_download/biological_unit/biological_unit_introduction.htm). The derived biological unit files (Biounit)—biological complexes that are based on the author's indications—are downloadable at the RCSB website. The Biounit files contain a MODEL record, as the NMR structures, when the original chain is duplicated to form the complex. Along with the Biounit coordinate file, we used the uniform PDB archive from the Data Uniformity Project to extract the 'curated' protein/sequence information (Westbrook *et al*., 2002). The mmCIF data files result from the reprocessing of PDB structures already present in the PDB database (ftp://ftp.rcsb.org/pub/pdb/data/structures/all/mmCIF). The differences between PDB files and mmCIF files concern the format, the nomenclature and the sequence structure consistency. The information contained in mmCIF files can be extracted using the CIF parse programs provided at the RCSB site.

3 BUILDING THE DATABASE

Programs were developed to automatically exclude undesirable complexes, characterize entries, chains and pairwise complexes by several attributes and extract representatives from the pool of complexes (Figure 1). A pairwise complex is defined as a binary combination of two chains present in the same 3D structure. In case of a higher multimeric state, corresponding annotation is added as well as indication of alternative binding modes. Only the structures solved by X-ray diffraction are included. The chains also must have the minimal length of 30 residues. A chain from the original PDB file can be repeated several times in a Biounit file. For example, PDB entry 1b0x is an eph receptor sam domain that reveals a mechanism for modular dimerization. The original PDB content (a.s.u.) contains only one chain A that has to be duplicated into the Biounit file to generate the biological complex (chain A-Model 1 and chain A-Model 2). Thus, the unique identifier for a particular Biounit chain in our database is a combination of original PDB id, original Chain id and Biounit MODEL id. In few PDB cases, we were not able to match original PDB chain(s) with the Biounit chain(s). These PDB are removed from the database and are listed in the link Info/List of excluded PDB structures/Excluded Biounit Files as 'unknown chain.' Since the 3D structure attributes are usually referenced by the PDB code and the original chain name, a Biounit chain (sometimes associated with the MODEL section number) has to be connected with the original name [e.g. to obtain the unique NCBI's GenInfo GI by using the SeqHound database (Michalickova *et al*., 2002)]. In the present work, chains have protein attributes such as the accession number in a sequence database (Swiss-Prot, EMBL, TrEMBL, etc.), keywords, SCOP classification (Hubbard *et al*., 1997), aminoacid sequence (SEQRES section of the PDB file and the 'genetic' domain sequence obtained from the ASTRAL compendium (Brenner *et al*., 2000); ATOM/HETATM sequence will be also provided in the future database update), and the numbering scheme of the protein segment in the sequence database, which does not match systematically the DBREF numbering scheme of the structure file. Additionally, each structure is associated with the name of the experiment, the resolution, the multimeric state [the number of chains that interact at least with one other chain in the Biounit complex with a mean ASA (solvent-accessible surface area) buried per chain > 250 Å^2] and the AEROSPACI score—an estimate of the quality of the structure obtained from the ASTRAL compendium.

A pairwise complex is defined by the names of the involved chains (including the MODEL section number) associated with their original chain names. For example, PDB entry 1b0x contains one pairwise complex in the Biounit file: chain A-Model 1 (original PDB chain A) interacts with chain A-Model 2 (original PDB chain A). A pairwise complex is classified HOMO if chains in the same PDB entry share >70% of sequence identity and BLAST *E*-value < 0.0001. About 75% of our database consists of HOMO pairwise complexes. The interface is characterized by the mean accessible surface area buried by each chain, computed by NSC program (Eisenhaber and Argos, 1993) and by the number of interface residues. The presence of a ligand, DNA or RNA, at the interface (≤5 Å of interface residues) or the existence of a disulfide bridge between chains is annotated. If the sequence database identifies a segment as a transmembrane one, then the pairwise complex is classified as 'membrane associated'.

Homo-n-ary and Hetero-n-ary annotations may occur when the multimeric state is higher than 2. In such complexes, all chains must interact with each other. For this purpose, we use the DBREF record extracted from the mmCIF file. We check whether two chains have the same DBREF (HOMO, if not—HETERO). Thus, if the DBREF record is missing, then the annotation is 'Not Determined' (2891 missing DBREF in 51506 Biounit chains in the database). An alternative binding mode means that a chain/protein may bind another chain/protein at more than one position (the DBREF record is also required). For example, Biounit entry 1f51 contains four chains, A

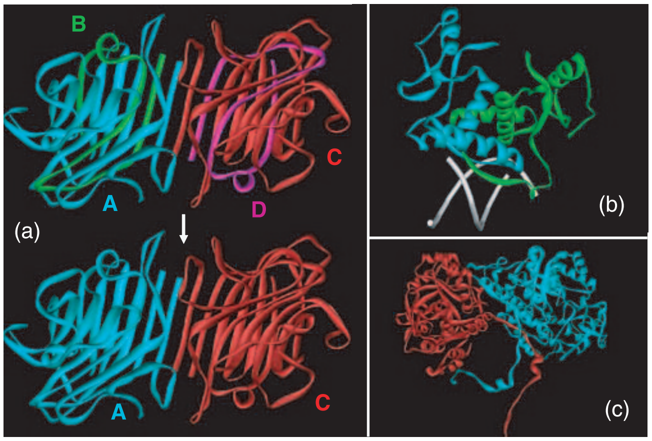


Fig. 2. Three types of illegitimate complexes are automatically detected: (**a**) interwoven chains—2ltn (AB and CD), (**b**) tangled chains—1cma AB and (**c**) disordered at the interface chains—1fcbAB.

(Sporulation response regulatory protein Spo0B, Swiss-Prot number P06535), B (Sporulation response regulatory protein Spo0B, Swiss-Prot number P06535), E (Sporulation response regulator Spo0F, Swiss-Prot number P06628) and F (Sporulation response regulator Spo0F, Swiss-Prot number P06628). Chains A, B and E interact with each other. With regards to A and B, chain E is a different protein. The complex A–B–E is annotated hetero-n-ary complex. Additionally, this complex presents alternate binding modes. For example, chain E interacts with chain A and B at different locations. Therefore, the sporulation response regulator Spo0F (chain E) contains two available binding sites for the sporulation response regulatory protein Spo0B (chains A and B). In this case, we preferred to detect homologies based on the DBREF record instead of the sequence identity. Thus, homo-n-ary pairwise complexes systematically have the sequence identity of 100%.

Three types of ‘illegitimate’ complexes are also detected and annotated: interwoven chains, tangled chains and termini parts of chains that interact but are disordered at the interface. Interwoven chains are identified by information in the DBREF record. Two chains are interwoven when two PDB chains are used to represent a single polymer with a residue gap. Generally, such sequences have to be consolidated into a single PDB chain (e.g. 2ltnAB and CD, 1cauAB, 1fmd1234). We found that 376 PDB entries contain such characteristics. We preferred to exclude such cases from our Biounit database even if some merged chains still interact with another chain(s) to form a complex (Figure 2a).

Pairwise complexes are marked ‘tangled’ when a free and unfolded segment of one chain interacts with another chain (>6 residues with ASA $\geq$ 40 Å each that interact exclusively with the second chain). The program can also identify some interwoven chains not identified in the first analysis (e.g. 1lgbAB or 1loaGH that do not have proper DBREF record). The algorithm is not perfect since some false positive cases were retrieved (e.g. 1ath). However, in this part certain trade-offs seem to be inevitable. Currently 752 ‘tangled’ pairwise complexes (369 PDB entries) in the Biounit database have been identified (e.g. 1cmaAB, 1parAB, see Figure 2b).

Table 1. Summary of the content of the DOCKGROUND database based on the Biounit data

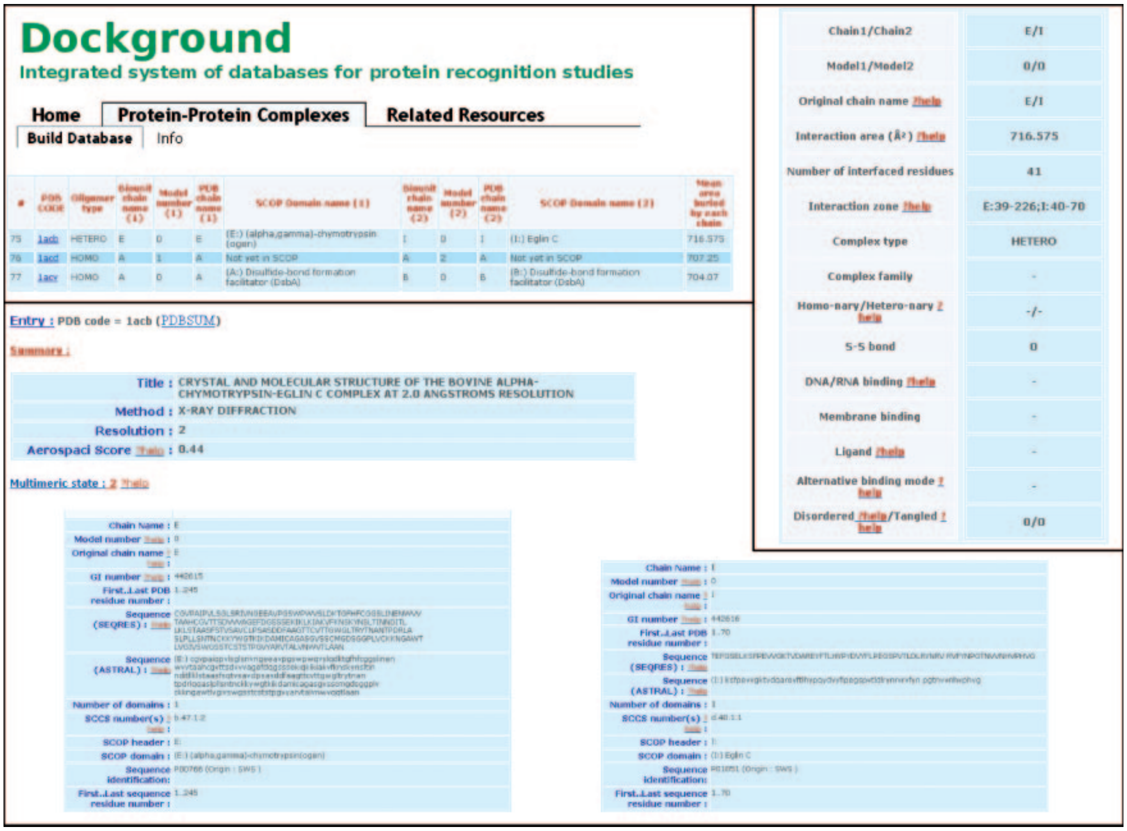
14 893 PDB entries
8628 (58%) dimeric complexes (multimeric state = 2 and a mean ASA buried by each chain $\geq$250 Å^2)
38 690 original chains (original PDB content)
51 506 biounit chains (original chain name associated with MODEL number)
36 111 biounit chains have SCOP SCCS number
26 025 biounit chains have 1 SCOP domain
8347 biounit chains have 2 SCOP domains
1271 biounit chains have 3 SCOP domains
236 biounit chains have 4 SCOP domains
141 biounit chains have 5 SCOP domains
67 biounit chains have 6 SCOP domains
6 biounit chains have 7 SCOP domains
18 biounit chains have 8 SCOP domains
67 220 pairwise complexes
50 132 (75 %) are HOMO pairwise complexes (BLAST: $\geq$70% of sequence identity between chains and *E*-value < 0.0001)
53 659 pairwise complexes (80%) possess an area $\geq$250 Å^2

The third illegitimate complex type involves chains with interacting unfolded termini parts (>10 residues with ASA $\geq$ 40 Å^2 each that interact with a similar segment of the other chain). We identified 203 pairwise complexes (83 PDB entries) with such characteristics (e.g. 1fcbAB, Figure 2c). The above attributes are stored in a relational database (implemented in PostgreSQL), which allows an efficient manipulation of the data. A form allows the user to view the data by requesting the PDB chain and pairwise complex table (Table 1 and Figure 3). Once the user’s input is completed, the server creates HTML pages for scrolling the PDB entry list and, for each PDB entry, the associated chains and pairwise complexes along with some of their attributes. The resulting page also offers an option to download a more comprehensive list of attributes (text file readable by Excel) as well as to create a representative list that will be sent to the user by Email.

4 SELECTION OF REPRESENTATIVE STRUCTURES

Working with representative structures allows one to avoid overrepresentation of some classes of proteins and a subsequent bias in results. For this purpose, we implemented a dynamic selection of a non-redundant subset by two different criteria: sequence identity and structure similarity. In two pairwise complexes, we allow a chain of one complex to be similar to a chain of another complex if the other chains are not similar. In a family of such complexes, we select representatives by the crystallographic resolution or the AEROSPACI score. Both lists are offered to the user along with a downloadable text file of hits attributes.

Several websites provide lists of PDB chains that are related by less than some fixed percentage of sequence identity. However some of these lists are apparently no longer maintained. The currently maintained lists are PISCES (Wang and Dunbrack, 2003) and PDB-REPRDB (Noguchi and Akiyama, 2003). PISCES is a public server for culling sets of protein sequences from the PDB by sequence identity. The database is weekly updated, the



Dockground

Integrated system of databases for protein recognition studies

Home | Protein-Protein Complexes | Related Resources

Build Database | Info

#	PDB CODE	Oligomer type	Biounit chain name (1)	Model number (1)	PDB chain name (1)	SCOP Domain name (1)	Biounit chain name (2)	Model number (2)	PDB chain name (2)	SCOP Domain name (2)	Mean area buried by each chain
75	1acb	HETERO	E	0	E	(E:) (alpha,gamma)-chymotrypsin (ogen)	I	0	I	(I:) Eglin C	716.575
76	1acd	HOMO	A	1	A	Not yet in SCOP	A	2	A	Not yet in SCOP	707.25
77	1acv	HOMO	A	0	A	(A:) Disulfide-bond formation facilitator (DsbA)	B	0	B	(B:) Disulfide-bond formation facilitator (DsbA)	704.07

Chain1/Chain2	E/I
Model1/Model2	0/0
Original chain name ?help	E/I
Interaction area (Å²) ?help	716.575
Number of interfaced residues	41
Interaction zone ?help	E:39-226;I:40-70
Complex type	HETERO
Complex family	-
Homo-nary/Hetero-nary ?help	-/-
S-S bond	0
DNA/RNA binding ?help	-
Membrane binding	-
Ligand ?help	-
Alternative binding mode ?help	-
Disordered ?help/Tangled ?help	0/0

Entry : PDB code = 1acb (PDBSUM)

Summary :

Title :	CRYSTAL AND MOLECULAR STRUCTURE OF THE BOVINE ALPHA-CHYMOTRYPSIN-EGLIN C COMPLEX AT 2.0 ANGSTROMS RESOLUTION
Method :	X-RAY DIFFRACTION
Resolution :	2
Aerospaci Score ?help :	0.44

Multimeric state : 2 ?help

Chain Name :	E
Model number ?help :	0
Original chain name ?help :	E
GI number ?help :	442615
First..Last PDB residue number :	1..245
Sequence (SEQRES) : ?help	[illegible]
Sequence (ASTRAL) : ?help	(E:) [illegible]
Number of domains :	1
SCCS number(s) ?help :	b.47.1.2
SCOP header :	E:
SCOP domain :	(E:) (alpha,gamma)-chymotrypsin(ogen)
Sequence identification:	P00766 (Origin : SWS)
First..Last sequence residue number :	1..245

Chain Name :	I
Model number ?help :	0
Original chain name ?help :	I
GI number ?help :	442616
First..Last PDB residue number :	1..70
Sequence (SEQRES) : ?help	[illegible]
Sequence (ASTRAL) : ?help	(I:) [illegible]
Number of domains :	1
SCCS number(s) ?help :	d.40.1.1
SCOP header :	I:
SCOP domain :	(I:) Eglin C
Sequence identification:	P01051 (Origin : SWS)
First..Last sequence residue number :	1..70

Fig. 3. The web view of a request producing a list of PDB entries (top left). PDB entry 1acb is a dimeric complex with only one interface between chains E and I and mean ASA buried by each chain 716 Å^2 (top right). Each involved chain is described by sequence and structure attributes (bottom).

sequences are extracted from mmCIF files and the user lists of PDB chains are processed by the server. For our purpose, we use the downloadable standalone package (http://dunbrack.fccc.edu/Guoli/pisces_download.php). The method uses PSI-BLAST alignments with position-specific substitution matrices derived from the non-redundant protein sequence database. Our choice of PISCES is based on an assumption that PSI-BLAST provides better estimates of sequence identity at longer evolutionary distances than the Needleman–Wunsch global alignment performed by PDB-REPRDB.

The structural classification is carried out using SCCS number from SCOP database. The SCCS number allows four types of clustering: class, fold, superfamily and family. Analysis of the Biounit database shows that out of 67 220 pairwise complexes 4479 are annotated as legitimate after removal of obsolete, interwoven, tangled and disordered complexes, specific cases (ligand, associated DNA or RNA at the interface, disulfide bridge, associated transmembrane segment) and those with the total mean ASA buried by each chain <250 Å^2 and the multimeric state higher than 2 (we chose to work with dimers). Two representatives selection modes are available: the pairwise mode, in which pairwise complexes representatives are selected from the previously filtered complexes (4479) and the oligomer mode, in which additional pairwise complexes associated with selected PDB entries are included (here 92). The additional pairwise complexes do not satisfy the previous filtering but this mode takes account for the whole Biounit file configuration. In the pairwise mode, 1476 representatives (1476 PDB entries) are selected based on the sequence identity <30% (960, when using the SCOP family level). Pairwise homodimer complexes represent 82% of the representative set. On the HTML page, homo and hetero complexes are separated to better visualize the results. In the oligomer mode, 92 pairwise complexes have been added. The representative set at 30% of sequence identity contains 1575 pairwise complexes (1488 PDB entries). Among them, 1460 pairwise complexes (1199 homo and 261 hetero PDB entries) contain only two chains that interact with a mean ASA buried per chain >250 Å^2 (only one interface: 'true' dimers). In the same set, 10 PDB entries contain 2 interfaces, 11 entries 3 interfaces, 1 entry 4 interfaces, 3 entries 5 interfaces, 1 entry 8 interfaces, 1 entry 16 interfaces, and 1 entry 19 interfaces.

Finally, the 'easy mode' allows users to access a precompiled dataset of representative complexes at 30% sequence identity based on the best resolution. It contains true dimeric complexes (currently, 1460 PDB entries) obtained by the oligomer mode.

Nevertheless, caution should be exercised in transferring the oligomeric state of a complex to other members of the protein family. Indeed, some examples having a high or near identical homology show a different complex configuration. For example, in the case of protein LicT mutations which occur on key functional residues provoke massive tertiary and quaternary rearrangements (PDB entry 1TLV). Such mutations are sometimes required to crystallize active (or inactive) form of the protein.

Table 2. Summary of the comparison of complexes contained in the original PDB dataset, Biounit dataset and PQS dataset

	Original PDB 8840 PDB entries (33 994 pairs; 30 523 chains)	Biounit 8728 PDB entries (39 511 pairs; 29 902 chains)	PQS 9909 PDB entries (44 836 pairs; 34 509 chains)
Mean ASA/chain (0; 1000) Å^2	14 298 pairs	14 303 pairs	16 979 pairs
Mean ASA/chain (1000; 5500) Å^2	18 241 pairs	22 814 pairs	23 168 pairs
Mean ASA/chain >5500 Å^2	1455 pairs	2394 pairs	4689 pairs
Dimer	4776 entries (54%)	4983 PDB (57%)	5584 PDB (56%)
Average multimeric state[a] (if states >14 are removed)	3.25	3.14	3.23
Average number of interfaces per chain	1.11	1.32	1.29
Average number of interfaces per PDB entry	3.84	4.52	4.52

The multimeric state is the number of chains that interact with at least one other chain, with mean ASA buried by chain >250 Å^2. Pair means a pairwise complex.
[a]Value for the Vakser's 1997 database, 2.79.

5 COMPARISON OF BIOUNIT AND PQS QUATERNARY CONFIGURATION

It is now acknowledged that an interface >1000 Å^2 is likely to be biological; however, this is still an approximation (Carugo and Argos, 1997; Dasgupta *et al*., 1997; Janin 1997; Janin and Rodier, 1995). Currently there is no accurate method to discriminate the biological interface from the crystal-packing one (Bahadur *et al*., 2004). The PDB provides access to putative biological complexes, called Biounit, which are based on the author's indications. On the other hand, the Protein Quaternary Structure file (PQS) server is an internet resource that makes available coordinates for probable quaternary states for crystallographically-determined structures in the PDB [http://pqs.ebi.ac.uk; (Henrick and Thornton, 1998)]. The predicted quaternary state is generated differently than in Biounit. We quantified the output of these two existing sources of biological complexes and compared the results to the original PDB content. However, it is important to emphasize that the results involving the original PDB content were expected because, as mentioned in Section 2, it contains the asymmetric unit and not the biologically functional unit.

The original PDB content was extracted from the mmCIF files. Such files may contain monomers, biological complexes and crystal-packing complexes. In PQS an automatic procedure is used to generate putative biological complexes. The complexes are built by progressive addition of monomeric chains that are considered to contribute to the assembly. The procedure is recursive allowing detection of quaternary structures where the contents of the asymmetric unit are not in contact with all other symmetry-related members of the final assembly. An automatic discrimination of potential quaternary structures between crystal-packing and biological oligomers is performed using an empirical score.

We compared the source files including the PDB entries. The PDB entries were limited to those deposited after January 1, 1999 (for the earlier entries, the Biounit data are based not only on the information provided by the depositor but also on supporting information obtained from the Swiss-Prot or PQS databases). The PDB entry list has been created with the entries.idx file at ftp://ftp.rcsb.org/pub/pdb/derived_data. Among the 29 327 PDB entries (January 26, 2005), 16 343 entries were selected (determined by X-ray diffraction, not obsolete, deposited in or after 1999, with the Biounit and the PQS structure file in PDB format). We discarded 341 entries containing non-protein molecules, as well as 119 proteins associated with high multimeric states. The analysis of complexes was performed on the 15 883 remaining entries. Only chains with ≥30 residues that interact with another chain were included. Thus, in total, 11 652 PDB entries are at least in one dataset as a complex and 6797 PDB entries are in all three datasets (Table 2).

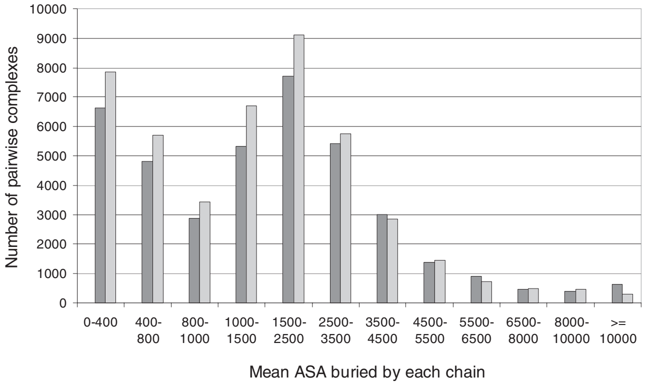


Fig. 4. Distribution of pairwise complexes. Histogram shows the frequency of mean area range buried by each chain in complexes in the original PDB dataset (stripes), Biounit dataset (dark gray) and PQS dataset (light gray).

5.1 Interface area

The interface is characterized by the mean ASA (solvent-accessible surface area) buried by each chain. The shape of the distribution is similar in the three datasets (Figure 4). However, the number of pairwise complexes that have at least 800 Å^2 buried area per chain is larger in PQS and Biounit datasets than in the original PDB content. As expected, Biounit and PQS datasets contain more probable or true quaternary complexes, involving more interfaces, than the original PDB content. The number of pairwise complexes that have 1000–3500 Å^2 buried area per chain is significantly larger than in the other ranges, except the 0–800 Å^2 range. These values are in agreement with the ones observed in confirmed biological complexes (Jones and Thornton, 1996). An important factor is the number of 0–800 Å^2 buried areas per complex. We found that 52% of PDB entries in the Biounit dataset have a multimeric state higher

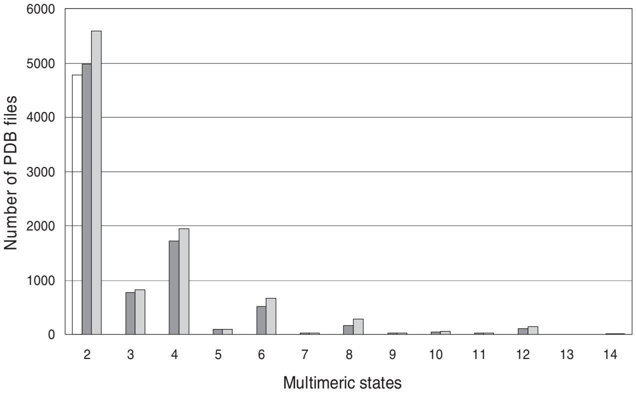


Fig. 5. Distribution of PDB entries. Histogram shows the frequency of multimeric state in PDB file in the original PDB dataset (stripes), Biounit dataset (dark gray) and PQS dataset (light gray). The multimeric state is the number of chains that interact with at least one other chain, with mean ASA buried by chain >250 Å^2.

than two, along with at least one buried area >800 Å^2 (41% in the original PDB content). This indicates that the 0–800 Å^2 range of buried area is occupied by 'secondary' (smaller) interfaces. Finally, PQS dataset tends to have the largest number of complexes in most ranges (except 3500–4500, 5500–6500 and >10000 Å^2).

5.2 Multimeric state

The multimeric state is defined by the number of chains that interact at least with one other chain, with the interface area ≥500 Å^2 (mean buried area per chain 250 Å^2). The shape of the multimeric state distribution is similar in each dataset and shares the same feature: the number of entries in the 'odd-meric' state is smaller than in the subsequent 'even-meric' state (Figure 5). The distribution also clearly shows that the dimeric state is the most occupied one: 54–57% in any dataset. In the first 14 multimeric states, the average multimeric state (average number of interacting chains in a PDB entry) is >3 in any dataset (Table 2) and is significantly higher than in the previous Vakser's dataset (2.79). The reason is that protein–protein co-crystallized complexes are now more commonly determined. The occupancy of higher multimeric states greater than the 14th contains <1% of the PDB entries, so we neglected them in the analysis. The average multimeric state is similar in each dataset. However, we detected a significant redistribution of PDB entries, caused by rebuilding of the oligomers in PQS and Biounit compared to the original PDB content. The average multimeric state in PQS is higher than in Biounit because of the larger occupancy of the most occupied states (states 2–12, with exception of states 5, 7, 11 and 13, which are the lowest occupancy states). However, the highest average multimeric state in the original PDB content (3.25) is not a consequence of the larger occupancy of multimeric states but rather a consequence of the lower occupancy of the dimeric and trimeric state. The average number of interfaces for one chain (Table 2) clearly shows a difference between the original content (1.11) and PQS and Biounit datasets (1.29 and 1.32, respectively). As expected, Biounit and PQS datasets contain more 'dense' complexes with more interfaces. Finally, the average number of interfaces for one PDB entry is 4.52 for PQS and Biounit (Table 2) and 3.84 for the original PDB content. This is also in agreement with more interacting complexes in PQS and Biounit datasets than in the original PDB content.

6 CONCLUSIONS AND FUTURE DIRECTIONS

The first release of the DOCKGROUND resource implements a comprehensive database of co-crystallized (bound–bound) protein–protein complexes, providing foundation for the upcoming expansion to unbound (experimental and simulated) protein–protein complexes, modeled protein–protein complexes and systematic sets of docking decoys. DOCKGROUND describes the interface of each pairwise complex in PDB entry by several attributes. The database is queryable by various descriptors, including AEROSPACI score (a global measure of the quality of the structure, assumed to be better than the resolution alone), a user-defined range of mean ASA buried per chain and the option to exclude various undesirable complexes (DNA/RNA or membrane associated, alternative binding modes and so on). DOCKGROUND allows selection of a representative list based on the user-defined percentage of sequence identity. DOCKGROUND is updated quarterly to reflect the growth of the PDB.

The current DOCKGROUND release contains additional features that allow users to submit a sequence to retrieve complexed homologs, therefore identifying putative partners and/or its quaternary state. In the future, corresponding components of DOCKGROUND will be integrated in the pipeline of the @TOME server (http://bioserv.cbs.cnrs.fr) to generate more precise models that take into account the quaternary environment (Douguet and Labesse, 2001).

An important aspect in designing databases of protein–protein complexes is the choice of the source of biological quaternary state. The original PDB content showed to be inappropriate. The difference between Biounit and PQS is at least 19% for shared PDB entries (37% for the 10 486 analyzed entries). A previous analysis performed by the authors of the PQS database showed that approximately one-third of their database is incorrect, one-third is correct and the last one-third have an unknown quaternary state (Henrick and Thornton, 1998). So far, such evaluation has not been performed on the Biounit data, which is the responsibility of the authors of deposited structures. Nevertheless, caution should be exercised in using this database too because discrepancies exist between the functional complex (e.g. disclaimed in the publication) and the Biounit one [example of the Lipid Transfer Protein (LTP) PDB1TUK not present in the Biounit database as a homodimeric functional protein]. However, additional criteria might be used to improve the quality of the database by applying the procedure developed by Bahadur *et al*. (2004), which showed the success rate of 93–95% on their complete homodimeric set (combination of the non-polar interface area and the fraction of buried interface atoms). In our study, we considered the benefit of the quantity of data to be more important than human errors (crystal-packing complexes annotated as biological ones).

The described resource is the first stage in DOCKGROUND development. Future development will include a better check of the validity of the source information, especially the sequence data (e.g. highlight potential mutations), advanced complex characterization (function, stability—obligate versus transient and so on), algorithms for simulating unbound structures from the co-crystallized components and the datasets of such structures, datasets of model–model complexes and docking decoys corresponding to all the protein complexes sets. The DOCKGROUND resource will improve our understanding of protein–protein interactions and will assist in developing better prediction tools.

ACKNOWLEDGEMENTS

The authors are grateful to Zhengwei Zhu for assistance with the DOCKGROUND website and Ying Gao for helpful suggestions. The study was supported by NIH grants R01 GM074255 and R01 GM61889. Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of Interest: none declared.

REFERENCES

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

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol*., **280**, 1–9.

Brenner,S.E. *et al*. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*., **28**, 254–256.

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

Dasgupta,S. *et al*. (1997) Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, **28**, 494–514.

Douguet,D. and Labesse,G. (2001) Easier threading through web-based comparisons and cross-validations. *Bioinformatics*, **17**, 752–753.

Eisenhaber,F. and Argos,P. (1993) *J. Comput. Chem*., **11**, 1272–1280.

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

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

Hubbard,T.J. *et al*. (1997) SCOP: a structural classification of proteins database. *Nucleic Acids Res*., **25**, 236–239.

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

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

Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Keskin,O. *et al*. (1998) Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci*., **7**, 2578–2586.

Keskin,O. *et al*. (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci*., **13**, 1043–1055.

Larsen,T.A. *et al*. (1998) Morphology of protein-protein interfaces. *Structure*, **6**, 421–427.

Lijnzaad,P. and Argos,P. (1997) Hydrophobic patches on protein subunit interfaces: charactersitics and prediction. *Proteins*, **28**, 333–343.

Lo Conte,L. *et al*. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol*., **285**, 2177–2198.

Lu,H. *et al*. (2003) Development of unified statistical potentials describing protein–protein interactions. *Biophys. J*., **84**, 1895–1901.

Marshall,G.R. and Vakser,I.A. (2005) Protein–protein docking methods. In Waksman,G. (ed.), *Proteomics and Protein–Protein Interaction: Biology, Chemistry, Bioinformatics, and Drug Design*. Springer, NY, pp. 115–146.

Michalickova,K. *et al*. (2002) SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics*, **3**, 32.

Noguchi,T. and Akiyama,Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res*., **31**, 492–493.

Papoian,G.A. and Wolynes,P.G. (2003) The physics and bioinformatics of binding and folding—an energy landscape perspective. *Biopolymers*, **68**, 333–349.

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

Russell,R.B. *et al*. (2004) A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol*., **14**, 313–324.

Tovchigrechko,A. and Vakser,I.A. (2001) How common is the funnel-like energy landscape in protein–protein interactions? *Protein Sci*., **10**, 1572–1583.

Tovchigrechko,A. *et al*. (2002) Docking of protein models. *Protein Sci*., **11**, 1888–1896.

Vajda,S. *et al*. (2002) Meeting report: modeling of protein interactions in genomes. *Proteins*, **47**, 444–446.

Vakser,I.A. *et al*. (1999) A systematic study of low-resolution recognition in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **96**, 8477–8482.

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

Westbrook,J. *et al*. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res*., **30**, 245–248.