

KBDOCK 2013: a spatial classification of 3D protein domain family interactions

Anisah W. Ghoorah¹, Marie-Dominique Devignes², Malika Smaïl-Tabbone¹ and David W. Ritchie^{3,*}

¹Université de Lorraine, LORIA, Campus Scientifique, BP 239, 54506 Villers-lès-Nancy, France, ²CNRS, LORIA, Campus Scientifique, BP 239, 54506 Villers-lès-Nancy, France and ³INRIA Nancy Grand Est, LORIA, Campus Scientifique, BP 239, 54506 Villers-lès-Nancy, France

Received August 15, 2013; Revised November 2, 2013; Accepted November 4, 2013

ABSTRACT

Comparing, classifying and modelling protein structural interactions can enrich our understanding of many biomolecular processes. This contribution describes Kbdock (<http://kbdock.loria.fr/>), a database system that combines the Pfam domain classification with coordinate data from the PDB to analyse and model 3D domain–domain interactions (DDIs). Kbdock can be queried using Pfam domain identifiers, protein sequences or 3D protein structures. For a given query domain or pair of domains, Kbdock retrieves and displays a non-redundant list of homologous DDIs or domain–peptide interactions in a common coordinate frame. Kbdock may also be used to search for and visualize interactions involving different, but structurally similar, Pfam families. Thus, structural DDI templates may be proposed even when there is little or no sequence similarity to the query domains.

INTRODUCTION

Many biological processes involve protein–protein interactions (PPIs). Thus, comparing and classifying PPIs can enrich our understanding of biology. To date, some 90 000 protein structures have been deposited in the Protein Data Bank (PDB) (1). However, recent analyses have shown that only about half of the expected number of human PPIs have so far been detected experimentally (2), and that only a small fraction of these have a 3D structure in the PDB (3). Nonetheless, there is a growing appreciation that template-based approaches are becoming increasingly useful for proteome-scale modelling of PPI networks (4). Furthermore, recent results from the CAPRI blind docking experiment and other studies have shown that using structural homology can significantly improve the quality of protein docking predictions (5,6).

There is, therefore, a need for user-friendly resources that can help to navigate networks of protein structure interactions and to propose templates for homology docking.

Here we present a major update to our Kbdock database system (7). Briefly, Kbdock combines sequence alignments from Pfam (8) with coordinate data from the PDB to classify the spatial arrangements of domain–domain interactions (DDIs) by Pfam family. The main feature that distinguishes Kbdock from other structural PPI or DDI databases [e.g. SCOPPI (9), SCOWLP (10), PiSite (11), 3DID (12), GWIDD (13), IBIS (14), ProtCID (15), InterEvol (16), PrePPI (17) and Interactome3D (18)] is that it uses a spatial clustering algorithm to define domain family binding sites (DFBSs). When all of the members of a Pfam domain family are superposed in a common coordinate frame, each DFBS describes the approximate spatial location of a cluster of binding sites without needing to enumerate individual domains or interaction residues. This allows existing DDIs to be described concisely at the Pfam family level using pairs of DFBSs, and it allows homology docking templates to be proposed by generating candidate model interfaces from the known binding sites within Pfam families. Furthermore, the 2013 version of Kbdock allows the user to search for DDIs across other structurally similar Pfam families using results from our ‘Kpax’ protein structure alignment algorithm (19). This allows more distantly related DDIs to be retrieved, which might have little or no sequence similarity to the query domains. Thus, Kbdock can propose structure-based homology docking templates, which might be difficult to find using only sequence-based similarity searches. We describe here an example of such a structural modelling scenario for the case of a TIM barrel bienzyme complex.

The current Kbdock database is built from the June 2013 snapshot of the PDB and the latest version of Pfam (release 27.0). It collects and classifies hetero and homo DDIs, as well as all domain–peptide interactions (DPIs). Overall, the new database contains 288 309

*To whom correspondence should be addressed. Tel: +33 3 83 59 30 45; Fax: +33 3 83 59 30 79; Email: dave.ritchie@inria.fr

domain structures belonging to 6516 Pfam entries and involving 239 494 DDIs and 11 852 various DPIs. Compared with the original version of Kbdock, which was built in 2009 and which stored only hetero DDIs, these figures correspond to an increase of ~50% over the previous number of structures and interactions.

Kbdock has an easy-to-use web interface, and all queries may be expressed using Pfam IDs or by providing the PDB codes or amino acid sequences of the domains of interest. The user may also provide a pair of query domains that are presumed to interact to search for similar DDIs or to define structural templates for docking. The results of queries against the database may be visualized in a common coordinate frame using the Jmol plug-in, and relationships between DDIs may be navigated visually using a Cytoscape plugin (20). The user may download the coordinate files of the superposed DDIs and a multiple sequence alignment in which the interaction residues are annotated. Thus, Kbdock provides a powerful and user-friendly interface to explore and visualize known DDIs and DPIs and to find knowledge-based templates with which to model unsolved protein complexes.

MATERIALS AND METHODS

Defining Pfam domain family binding sites

The Kbdock database is populated using a series of in-house scripts that (i) extract the protein chains for a PDB file, (ii) feed the chain sequences to PfamScan (21) to identify annotated Pfam domain and peptide families, (iii) cut each chain into separate domains and peptides, (iv) count the number of atomic contacts between each pair of domains or a domain and a peptide, (v) filter out identical copies of the same interaction using a sequence-based within-PDB filter and (vi) calculate the interface surface area of all domain–domain contacts using DSSP (22). We use the same criteria as Stein *et al.* (12) to define a DDI or a DPI (i.e. essentially 5 or more contacts are required for a physical interaction).

We then classify each DDI as ‘intra’ or ‘inter’ and ‘homo’ or ‘hetero’ according to whether the interaction is within one chain or across two chains, and whether the interaction involves the same or different chains, respectively. If an inter-chain hetero DDI has multiple crystal contacts, we assume that the biological interface is the one with the largest interface area (23,24). We retain all distinct inter-chain homo interactions because in this case it can be less clear how to identify a biologically relevant interaction (15). Next, we annotate every interface residue as ‘core’ or ‘rim’ depending on their solvent accessibility (25).

For every Pfam domain having structures found by the aforementioned protocol, the domain sequences are aligned using HMMER (26). This multiple sequence alignment is used to place all of the domains and their DDI partners in a common coordinate frame using the ProFit program (<http://bioinf.org.uk>). For each superposed DDI, a binding site ‘centre residue’ is calculated by selecting the core and rim residues that contact the partner domain and

by selecting the C_{α} atom that is nearest to a weighted average of the core (75%) and rim (25%) C_{α} coordinates. We then calculate a ‘binding site direction vector’ that points from the domain’s centre of mass to the centre residue’s C_{α} atom. Within each Pfam family, the angular distance between pairs of binding site vectors is then clustered using agglomerative hierarchical clustering with a threshold of 24° to define DFBSs. Finally, for each DFBS, we select a representative list of distinct Pfam partners to define a set of domain family interactions (DFIs). The same processing steps are used to identify and cluster DPIs.

It is worth emphasizing that the aforementioned algorithm defines and clusters binding sites, not interfaces. A binding site is defined by a patch of surface residues, and a DFBS is defined by a cluster of one or more surface patches located in approximately the same surface region of a group of superposed Pfam domains. The partner domains play no role in defining a binding site except to select the residues to be considered on the domain of interest. The use of a binding site direction vector allows us to avoid the difficult problem of how to compare in some more precise way the similarity of two binding site patches that might be made up of different types and arrangements of surface residues.

Although there is currently no accepted standard for how, precisely, to define a binding site or how to distinguish interfaces that share a certain number of residues or residue contacts, we find that, in practice, our approach works well to distinguish different binding sites from one another. However, because it deliberately does not take into account the specific residue contacts within individual interfaces, it can sometimes place into the same cluster binding sites that have similar direction vectors but that have only a few residues in common. We are considering whether it would be beneficial to distinguish such cases in a future version. In any case, it should be borne in mind that our definition of a DFBS derives from a simple spatial heuristic with a somewhat arbitrary angular threshold.

Defining Pfam structural neighbours

Although Pfam collects related families into some 515 ‘clans’ (27), currently only 4563 of 14 831 Pfam families belong to a clan, with the rest being unassigned. Thus, we cannot depend on Pfam clan annotations to provide structural neighbours for every family of interest. To circumvent this limitation, we have calculated and stored structural Pfam neighbours for every Pfam family using our Kpax structure alignment algorithm (19). In Kpax, the similarity between two protein chains is calculated as a sum of Gaussian overlaps between pairs of aligned C_{α} atoms. This score is then normalized using the geometric mean of the chain lengths. In our experience, a normalized Kpax score of ≥ 0.3 often denotes a significant alignment and superposition. To build a neighbour list, we first used Kpax to calculate a ‘centre’ structure for the domains within each Pfam family, and we then used Kpax again to calculate an all-versus-all similarity matrix between the centre structures. Sorting the rows of this matrix by Kpax

similarity score then gives a ranked list of Pfam family neighbours for each Pfam family, from which the top 1000 family neighbours for each Pfam are stored in the Kbdock database. Thus, when searching for DDIs that involve a given query domain, or when navigating related interactions, the 2013 version of Kbdock can rapidly retrieve similar DDIs using its pre-calculated neighbour lists.

We recently compared the performance of Kpax with the widely used TM-align (28) protein structure alignment program, and with the fast Yakusa (29) structure aligner and database search program. When searching the CATH database (version 3.4, 11330 domains at the 35% sequence identity level) using 213 different CATH domains as queries, we found that Kpax was faster and considerably more accurate than Yakusa, and that it gave almost the same high level of recall and precision as TM-align (Kpax gave an aggregate area under the curve of 0.966 compared with 0.976 for TM-align), while being over 100 times faster (19). Thus, we are confident that searching our Pfam structural neighbour lists provides a fast and reliable way to retrieve structural homologues of a given query domain.

Finding DDI homology templates

Given two query domain structures, Kbdock searches for DDIs involving the same Pfam families as the query domains. We call any DDIs that satisfy this search ‘full-homology’ (FH) templates. However, although we find that many Pfam domains have just one DFBS, some have multiple DFBSs. Therefore, if several different FH DDIs match the query domains, Kbdock outputs a proposed docking model for each distinct pair of binding sites.

On the other hand, if no FH templates are found, Kbdock searches for and outputs DDIs containing the individual query domains because these can still provide useful information for a docking calculation (30). We call such DDIs ‘semi-homology’ (SH) templates. In these cases, the query domain is superposed onto each template in turn to propose a binding site on the query domain. If several SH templates are found for a given query domain, Kbdock ranks them in order of sequence similarity to the query. Sequence similarity is calculated using either the Pfam consensus alignment or the Kpax structural alignment, as appropriate, and by summing the number of aligned residues belonging to the same chemical group (i.e. polar: S, T, N, Q, C, Y; acidic-polar: D, E; basic-polar: K, R, H; non-polar: G, A, V, L, I, P, M, F, W). When viewing templates generated from structural neighbours, the Kbdock results page includes details of the number of structurally aligned residues and the corresponding RMSD. If desired, the user may launch a protein docking run using the *Hex* server (31), with the retrieved templates being passed directly from the Kbdock results page.

Finding DDI templates using Pfam structural neighbours

When no FH templates exist for a given pair of query structures, or when only SH binding sites are found, the

user might still wish to consider more remote homologous interactions that could exist between members of other Pfam families. Therefore, the Kbdock interface may be used to search for DDIs between structurally similar Pfam domains using the pre-calculated Pfam neighbour lists. However, because the retrieved neighbour family DDIs often require close visual inspection by the user, and because multiple such interactions might need to be considered, Kbdock does not provide a direct link to the *Hex* server for these cases.

Kbdock database implementation and web interface

A full description of the relational data model and schema diagram is provided under the Kbdock home page. The physical database is implemented using the MySQL database engine (<http://www.mysql.com>). Operations such as parsing and processing the data are implemented in the C and Prolog programming languages. Binding site direction vectors are clustered using R scripts (<http://www.r-project.org/>). The public Kbdock website and database files are available at <http://kbdock.loria.fr/>.

The Kbdock web interface is written mainly in the PHP scripting language (<http://php.net>). Some queries are processed using Prolog and Linux shell scripts. The Jmol plug-in is used for 3D visualization (<http://www.jmol.org>), whereas the Cytoscape plugin (<http://cytoscapeweb.cytoscape.org>) may be used to navigate the DDIs involving a given domain of interest. The web interface has been tested using several popular browsers for the Windows, Linux and Mac OS X operating systems. Scripts for creating high-resolution graphics locally using VMD (32) are available for download.

RESULTS

Kbdock database content

Table 1 summarizes the total numbers of non-redundant DDIs and DPIs stored in Kbdock. As noted in Methods, Kbdock retains only one ‘biological’ interaction for each pair of inter-chain hetero DFBSs, whereas it stores all distinct pairs of DBFSs for inter-chain homo interactions. This contributes to the relatively large number of stored inter-chain homo DDIs. It is also worth noting that Kbdock follows the Pfam convention of distinguishing between ‘annotated’ and ‘unannotated’ DPIs. The final three columns of Table 1 show that, after spatial clustering

Table 1. The total numbers of non-redundant DDIs and DPIs (second column) and family-level interactions (third to final columns) in Kbdock

Interaction type	Total	PFAMs	DFBSs	DFIs
Inter-chain hetero DDIs	20 126	2153	4001	2517
Inter-chain homo DDIs	128 019	3982	12 498	4433
Intra-chain hetero DDIs	21 134	2018	3021	1487
Intra-chain homo DDIs	3489	354	745	354
Annotated inter-chain DPIs	297	32	38	32
Annotated intra-chain DPIs	342	19	22	19
Unannotated DPIs	11 852	873	1341	—

by DBFSs, the PDB contains only a few thousand binding sites and interaction types when considered at the domain family level. A more detailed break down is provided on the Kbdock website.

Analysing DDIs and DPIs by Pfam family

To analyse the binding sites of a given Pfam family, the user may use the Kbdock 'Search' page to enter a Pfam identifier (e.g. Kunitz_legume), a Pfam accession number (e.g. PF00197), a keyword (e.g. inhibitor), an amino acid sequence or a PDB file of a protein structure. If a sequence or a structure is entered, the PfamScan utility is used to determine the Pfam accession number. Otherwise, the accession number is found directly from the Kbdock database. Kbdock then retrieves a non-redundant list of DDIs involving the query domain, grouped by their binding site. Figure 1 shows the hetero interactions found when Kbdock is queried using the Kunitz_legume protease inhibitor domain (PF00197). The Jmol plugin shows the retrieved DDIs in the coordinate frame of the query domain. The query domain is shown in grey and interacting residues are shown as wire sticks. The user may choose to view the DDIs together or individually. An annotated Pfam consensus alignment of the retrieved domains is also provided, in which each sequence is colour-coded according to the core, rim and centre residue assignments. Links to download the multiple sequence alignment and the superposed PDB files as a single compressed file are also available.

In a similar way, the user may analyse DPIs involving a given domain family. For example, Figure 2 shows examples of the DPIs belonging to three selected domain families. This figure was generated using VMD for clarity, whereas all graphics on the Kbdock web pages are drawn using Jmol for better portability. However, VMD scripts are provided, which can render each Jmol scene locally in high resolution.

Retrieving docking templates

To find docking templates, the user enters two PDB codes or uploads two PDB files and he then specifies which pair of Pfam domains in those structures should be used as queries. If Kbdock finds one or more FH DDI templates for the query domains, it shows the superposed query and FH template(s) using Jmol along with colour-coded sequence alignments of the query and template domains showing the core, rim and centre binding site residues (similar to Figure 1). As before, the user may download the query and template structures in the superposed orientations. If no FH DDIs exist in the database, Kbdock will output a non-redundant list of SH templates with annotated sequence alignments along with a Jmol view of the superpositions.

It is also possible to use Pfam structural neighbours to find FH templates. To give a particular example, let us suppose we wish to model the complex between a glutamine-dependent amido transferase (GATase) and the cyclase domain of an imidazole glycerol phosphate (ImGP) synthase. The delivery of an amino group by the GATase to the cyclase site at the N-terminal region of the

KBDock
a resource for knowledge-based protein docking

Your query Pfam family is Kunitz_legume (PF00197)

Jump to >

Representative inter-chain hetero domain-domain interactions for Kunitz_legume (PF00197)

Show All inter-chain hetero domain-domain interactions

Query Family Kunitz_legume (PF00197)					Partner family				
Site ID	PDB	Pfam ID	Chain	Start End	Pfam ID	Chain	Start End		
PF00197_1_3e8l	Kunitz_legume	C	2	175	Trypsin	B	16 231		
PF00197_1_3bx1	Kunitz_legume	D	5	177	Peptidase_S8	B	27 274		
PF00197_2_3veq	Kunitz_legume	A	607	776	Trypsin	B	16 231		
PF00197_3_lava	Kunitz_legume	D	5	177	Alpha-amylase	B	19 322		
PF00197_4_2lwt	Kunitz_legume	B	5	177	Thioredoxin	A	21 114		

Jump to >

Superposition for Kunitz_legume (PF00197)

Select interaction

- All
- 3e8l_C_2_175
- 3bx1_D_5_177
- 3veq_A_607_776
- lava_D_5_177
- 2lwt_B_5_177

Select binding site

- Site_1
- Site_2
- Site_3
- Site_4

Jmol, 5

Jump to >

Pfam consensus sequence alignment for Kunitz_legume (PF00197) with binding site information

Note (i) Top sequence is the Pfam family consensus sequence
(ii) Binding site residue color-coding scheme: "center", "rim", "rim"
(iii) Position cursor on a residue to display the PDB residue label
(iv) Table of Pfam Consensus Amino Acid Classes

```

Consensus_Pf00197      11DS-6S...1c...s66.0Y...YLPLhucGG...1LAPISGHESCP
PF00197_1ava_D_5_177   PNHDTGHE-LR-ADA-HY...YVLSANRAGGG...LTMFQGHGRPC
PF00197_2lwt_B_5_177   PNHDTGHE-LR-ADA-HY...YVLSANRAGGG...LTMFQGHGRPC
PF00197_3bx1_D_5_177   PNHDTGHE-LR-ADA-HY...YVLSANRAGGG...LTMFQGHGRPC
PF00197_3e8l_C_2_175   PNVSDGDA-VQLHLGG-VPLTTSQALDFRGG...1ETLND--ACK
PF00197_3veq_A_607_776 -LVDAEGLN-VE--HGG-TY...YVLPILHAGGG...1ETAKTNEPCP

Consensus_Pf00197      LQVVPpPs-1s.cchPIRTSS.h.psta1scu...oh1slfussssCs
PF00197_1ava_D_5_177   LVSQDQHQH-DGPPVITPVG--VAPSDKIRLSTVDISFRATTCCL
PF00197_2lwt_B_5_177   LVSQDQHQH-DGPPVITPVG--VAPSDKIRLSTVDISFRATTCCL
PF00197_3bx1_D_5_177   LVSQDQHQH-DGPPVITPVG--VAPSDKIRLSTVDISFRATTCCL
PF00197_3e8l_C_2_175   SYVGEAE-1D-HGLVGSKA---SATSPQWGLGRVYSFSDHPV1
PF00197_3veq_A_607_776 LTVVRSPNEVS-KGEPIRTSSQ--RSLFIPRG---SLVALGFANPSCA

Consensus_Pf00197      sss...hwoVscp.t...G...svKluppc...tt.hsfFHEKS
PF00197_1ava_D_5_177   GST-----EHWDSLEAA-GRR--HVITGPVKDP--SPSGREAFRIEY
PF00197_2lwt_B_5_177   GST-----EHWDSLEAA-GRR--HVITGPVKDP--SPSGREAFRIEY
PF00197_3bx1_D_5_177   GST-----EHWDSLEAA-GRR--HVITGPVKDP--SPSGREAFRIEY
PF00197_3e8l_C_2_175   GST-----AEIDKQETH-GQSGPPTAGDTY1---LWFSFSGARST
PF00197_3veq_A_607_776 ASP-----WTVVDSPPQ---GP---AWLSQKQL---PEHDLVFRFEV

Consensus_Pf00197      UC.s...tsyKLIAC.p.....pc.t.lgIph.psgSRRLVlocp.SP
PF00197_1ava_D_5_177   SG-AEVEHKLHSCG-----DMC-QQLQVFRD-LKGSANFLGATE-PY
PF00197_2lwt_B_5_177   SG-AEVEHKLHSCG-----DMC-QQLQVFRD-LKGSANFLGATE-PY
PF00197_3bx1_D_5_177   SG-AEVEHKLHSCG-----DMC-QQLQVFRD-LKGSANFLGATE-PY
PF00197_3e8l_C_2_175   EE--TQVYKLACSEFC-KIAC-PEVQ-SFV-VHRTLLGIGG--EN
PF00197_3veq_A_607_776 SH-SNIDVYKLLYQHDEE--DVKCDQYTGIRHD-RGNRRLVTEE-NP

Consensus_Pf00197      L.lhh.....
PF00197_1ava_D_5_177   HVVVFRK-----
PF00197_2lwt_B_5_177   HVVVFRK-----
PF00197_3bx1_D_5_177   HVVVFRK-----
PF00197_3e8l_C_2_175   FTVQFK-----
PF00197_3veq_A_607_776 LELV-L-----
  
```

Jump to >

Download data for Kunitz_legume (PF00197)

Contents	Interactions		Type
	ALL	REP	
Interface details of interactions involving query family Kunitz_legume (PF00197)	<input type="checkbox"/>	<input type="checkbox"/>	Txt
PDB files of interactions involving query family Kunitz_legume (PF00197)	<input type="checkbox"/>	<input type="checkbox"/>	Tar.gz
Consensus sequence alignment of query family Kunitz_legume (PF00197)	--	<input type="checkbox"/>	Pdf

Figure 1. A screenshot of the Kbdock results for the query domain family, Kunitz_legume. The results page consists of four sections: (A) a non-redundant list of hetero DDIs grouped by their binding site, (B) a Jmol view of the DDIs in the coordinate frame of the query domain (the query domain is shown in grey and interface residues are shown in wire-frame), (C) a Pfam consensus-based sequence alignment of the domains annotated with the core (green), rim (blue) and centre (red) binding site residues and (D) links to download the multiple sequence alignment and the superposed PDB files.

TIM barrel synthase is one of the steps in histidine biosynthesis (33).

The structure of a GATase/ImGP cyclase complex and the structures of the unbound domains have already been solved for the bacterial thermophile *Thermatoga maritima*

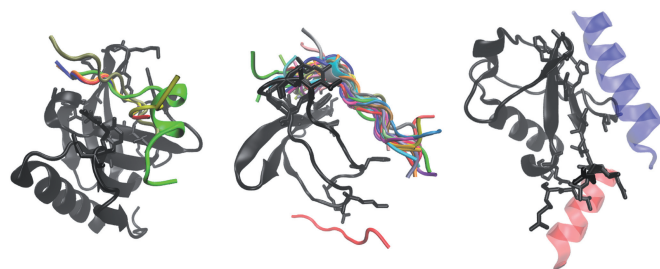


Figure 2. VMD views of superposed domain-peptide interactions for three different domain families. From left to right, the SspB family (PF04386) has six interactions involving one DFBS. The SH3_1 family (PF00018) has 40 interactions involving two DFBSs. Ubiquitin (PF00240) has two interactions, each using a distinct DFBS.

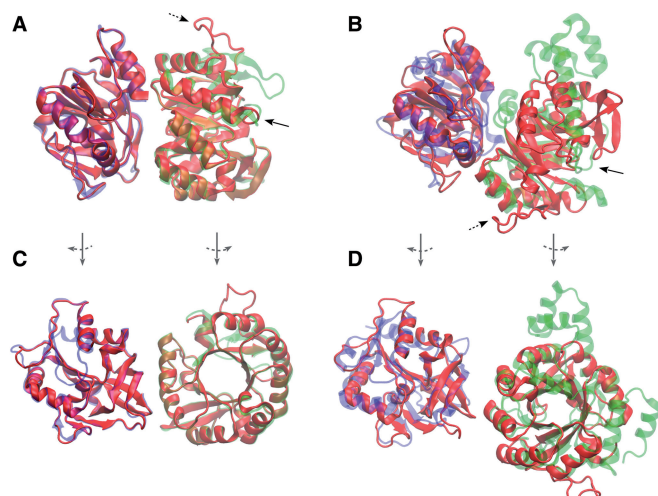


Figure 3. Homology modelling a TIM barrel bienzyme using Pfam structural neighbour interactions. (A) The superposition of the query domains in red (a GATase belonging to PF00117 from PDB code 1K9V; and an ImGP synthase, PF00977, PDB code 1THF) superposed onto the FH template shown in blue and green (PDB code 1GPW). All superpositions are calculated using Kpax. Both structures have a central TIM barrel co-linear with the intermolecular axis in the plane of the page. (B) The same query structures in red superposed onto the Pfam neighbour template (PDB code 2NV2) consisting of a SNO domain (PF01174) in blue and a SOR_SNZ domain (PF01680) in green. Solid arrows show the approximate location of the ImGP substrate binding site at the C-terminal ends of the β strands of the TIM barrel. Dashed arrows locate an exposed loop in the synthase query domain, which is also present in the FH template (1GPW), but which becomes helical in the neighbour template (2NV2). (C and D) The same structures but with each domain rotated away from its partner by 90° to give a view of the interface (now with the N to C direction of the TIM barrel going into the page). As can be seen, the 2NV2 structure also has a TIM barrel catalytic domain, although it is rotated by $\sim 180^\circ$ with respect to its homologue in the 1GPW structure. In other words, rotating the ImGP synthase and the SOR_SNZ domains by $\sim 180^\circ$ about the intermolecular axis would bring them into close register with the FH template.

(34). Hence, modelling this complex from the unbound domains should be trivially easy for any homology modelling protocol. Part A of Figure 3 shows VMD views of the FH template (PDB code 1GPW) that was retrieved when querying Kbdock using the GATase (PDB code 1K9V) and an ImGP synthase structure (PDB code 1THF) that belong to the GATase (PF00117) and His_biosynth (PF00977) Pfam families,

respectively. This shows the good structural overlay of the query and template domains, as expected.

Histidine biosynthesis is just one of the many enzymatic functions carried out by members of the TIM barrel family (35). However, finding more distant TIM barrel homologues can be challenging because their sequence similarities are often below the detectable level (35). To illustrate such a scenario, let us suppose that the 1GPW structure had not been solved. Part B of Figure 3 shows the template found when using the Kbdock structural neighbour list to find more remote DDI homologues. In this case, Kbdock considered the DDIs involving four neighbour Pfams of PF00117 and 16 neighbours of PF00977, and it found an interaction between the SNO (PF01174) and SOR_SNZ (PF01680) domains [a PLP synthase complex from *Bacillus subtilis* (36)] as the first instance of a candidate template (PDB code 2NV2) involving the two neighbour domains together. These domains have only 16.3 and 13.8% sequence identity with the query domains, respectively, and the PLP synthase structure carries some additional α -helices compared with the ImGP synthase.

It is worth noting that both GATase and SNO belong to Pfam clan Glutaminase_I (CL0014), and that the His_biosynth and SOR_SNZ cyclase domains both belong to the TIM_barrel clan (CL0036), although these relationships were not used explicitly. Figure 3 shows that the retrieved domains still have the same folds and interfaces as the query domains (Kpax structurally aligned 167 and 152 residues with RMSDs of 2.7 and 2.3Å between the cyclase and GATase domains, respectively), although one of the cyclase domains is slightly translated and is rotated by $\sim 180^\circ$ about the intermolecular axis with respect to its homologue.

Nevertheless, because the cyclase has the same orientation with respect to the GATase in both of these structures, the proposed neighbour template is consistent with the ammonia tunnelling model that has been proposed for the delivery of the amino moiety from GATase through the TIM barrel to the cyclase catalytic site (34). These observations confirm that our Pfam neighbour lists can help to find remote homologues. However, a template modelling protocol based on residue contacts would probably fail with this example because Figure 3 shows that it would be necessary to run a docking refinement calculation to correct the rotational and translational mismatch.

DISCUSSION

In recent years, many protein structure interaction databases have been described (37). For example, SCOPPI (9) classifies DDIs using geometric overlap and face angle scores of residue contact vectors (38). For a given SCOP family, SCOPPI outputs all PDB complexes involving the query. DDIs are grouped according to their partner domain. For each group, multiple sequence alignments annotated with the interacting residues are provided for both the query and partner family. SCOPPI also outputs its calculated interface type, area and volume, and the web interface provides a screen-shot of each interface and

external links to related publications. SCOWLP (10) is also a SCOP-based classification of protein domain and peptide interactions. For each SCOP family, SCOWLP performs pairwise structural alignments to identify common interacting residues with which to define and cluster 'binding regions' using a sequence-based similarity score. The user may browse the SCOP hierarchy to view interactions or to search for interactions involving a given SCOP family.

3DID (12) classifies DDIs and DPIs using hierarchical complete linkage clustering of groups of interface residues or 'interface profiles'. The first version of Kbdock was built from 3DID, although this dependency has since been removed. For a given Pfam family, 3DID outputs a list of its partner domains grouped by interface profile. ProtCID (15) also uses Pfam to assign a 'chain architecture', i.e. a list of Pfam identifiers, to each protein sequence in each PDB entry, and it clusters similar chain-chain interfaces using a distance-based pairwise amino acid similarity score. The ProtCID web interface supports queries by PDB ID or by pairs of protein sequences or Pfam identifiers, and it lists the PDB IDs of structures with matching chain architectures.

IBIS (14) stores experimentally determined and inferred physical interactions between proteins, peptides, DNA and RNA and other small molecules. IBIS defines DDIs using its manually curated 'CDD' domain definition and it classifies them using hierarchical complete linkage clustering of groups of interface residues. For a given query protein, IBIS outputs a list of its interaction partner proteins. The interactions are listed as DDIs, which are grouped by their partner domain and binding site. The identities of binding site residues on the query protein are also shown.

GWIDD (13) is a database of experimentally solved and predicted 3D structures of protein-protein complexes built from the PDB and the BIND (39) and DIP (40) databases. If GWIDD does not have a 3D structure for a given query, it builds a 3D model using Nest (41) to calculate structure-based superpositions. Interactome3D (18) predicts 3D structures of PPIs using the MODELLER (42) comparative modelling program. For a given UniProt query code, Interactome3D displays a PPI network in which each node leads to a page describing the protein, and each edge leads to a page describing the interaction. With a similar goal, PrePPI (17) aims to predict large-scale interaction networks between pairs of UniProt sequences using homology modelling followed by structural alignment searches against the PDB, and using evolutionary, functional and expression information to calculate Bayesian confidence scores.

Although all of these resources are useful and many of them provide lists of interaction residues and other derived quantities, most of them cannot be used to provide docking templates directly because they cannot be queried with two sequences or structures simultaneously. For example, only GWIDD, ProtCID, Interactome3D and PrePPI can process more than one sequence or structure at a time. Of these, Kbdock is somewhat similar to GWIDD and ProtCID. GWIDD can often produce a 3D model of a protein when it can

find a pairwise template with sufficient overall sequence similarity to the two queries. However, because GWIDD integrates PPIs from several sources, it contains a large number of modelled interactions, whereas Kbdock stores only experimentally solved structures. ProtCID also stores only experimentally solved structures and it can find PDB chains that match the chain architecture of the query structure(s). However, ProtCID operates only at the chain level, whereas Kbdock works at the domain level. Furthermore, neither GWIDD nor ProtCID allows the binding sites and common domain-level interactions of a given query to be visualized online.

More importantly, none of the aforementioned approaches explicitly attempt to classify or re-use the spatial arrangements between known protein binding partners. One of the early design aims for Kbdock was to help find candidate templates for protein docking, where the spatial orientations of pairs of interacting domains are of primary importance. Consequently, Kbdock has a rather unique DDI classification and template modelling pipeline: (i) it uses the Pfam consensus sequence to place all of the complexes involving a given Pfam domain family into a common coordinate frame; (ii) it uses the notion of 'core' and 'rim' interface residues to group the complexes by the spatial position of their binding site; (iii) it finds automatically the best DDI template with which to homology-model a complex of two given structures; (iv) if more than one interface is found, it proposes a model for each; (v) if no suitable FH template exists, it can still propose candidate SH binding sites for one or both interaction partners; and (vi) it uses lists of Pfam structural neighbours to propose candidate docking templates even when there is little or no sequence similarity with the query domains. Additionally, thanks to the Jmol plug-in, the Kbdock web server provides a convenient way to view and compare Pfam binding sites and calculated docking templates.

CONCLUSIONS

Kbdock provides a useful resource with a unique processing pipeline for analysing the 3D structures of DDIs within and between Pfam domain families and for finding knowledge-based docking templates to help predict the structures of unknown protein complexes. It also allows structural similarities between Pfam families to be exploited to find candidate docking templates even when no direct Pfam homologues exist. The 2013 version of Kbdock supports analyses of all available hetero and homo domain interactions and binding sites from the PDB, as well as domain-peptide interactions and binding sites. Online visualization is provided by the Jmol and Cytoscape plugins. All data and results are available for download in several formats, as well as VMD scripts for high-quality local visualization.

FUNDING

Agence Nationale de la Recherche [ANR-08-CEXC-017-01 (2009–2011) and ANR-11-MONU-006-02 (2011–2015)]. Funding for open access charge: INRIA.

Conflict of interest statement. None declared.

REFERENCES

- Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlic, A., Quesada, M. *et al.* (2013) The RCSB protein data bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
- Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.
- Stein, A., Mosca, R. and Aloy, P. (2011) Three-dimensional modeling of protein interactions and complexes is going omics. *Curr. Opin. Struct. Biol.*, **21**, 200–208.
- Vakser, I.A. (2013) Low-resolution structural modeling of protein interactome. *Curr. Opin. Struct. Biol.*, **23**, 198–205.
- van Dijk, A.D., Boelens, R. and Bonvin, A.M. (2005) Data-driven docking for the study of biomolecular complexes. *FEBS J.*, **272**, 293–312.
- Lensink, M.F. and Wodak, S.J. (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins*, **78**, 3073–3084.
- Ghoorah, A.W., Devignes, M.D., Smail-Tabbone, M. and Ritchie, D.W. (2011) Spatial clustering of protein binding sites for template based protein docking. *Bioinformatics*, **27**, 2820–2827.
- Punta, M., Coghill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
- Winter, C., Henschel, A., Kim, W.K. and Schroeder, M. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.
- Teyra, J., Paszkowski-Rogacz, M., Anders, G. and Pisabarro, T.M. (2008) SCOWLP classification: structural comparison and analysis of protein binding regions. *BMC Bioinformatics*, **9**, 9.
- Higurashi, M., Ishida, T. and Kinoshita, K. (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, **37**, D360–D364.
- Stein, A., Ceol, A. and Aloy, P. (2010) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
- Kundrotas, P.J., Zhu, Z.W. and Vakser, I.A. (2010) GWIDD: genome-wide protein docking database. *Nucleic Acids Res.*, **38**, D513–D517.
- Shoemaker, B.A., Zhang, D., Tyagi, M., Thangudu, R.R., Fong, J.H., Marchler-Bauer, A., Bryant, S.H., Madej, T. and Panchenko, A.R. (2012) IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.*, **40**, D834–D840.
- Xu, Q. and Dunbrack, R.L. (2011) The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.*, **39**, D761–D770.
- Faure, G., Andreani, J. and Guerois, R. (2012) InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.*, **40**, D847–D856.
- Zhang, Q.C., Petrey, D., Garzon, J.I., Deng, L. and Honig, B. (2013) PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.*, **41**, D828–D833.
- Mosca, R., Ceol, A. and Aloy, P. (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.
- Ritchie, D.W., Ghoorah, A.W., Mavridis, L. and Venkatraman, V. (2012) Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, **28**, 3274–3281.
- Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Lotia, S., Pico, A.R., Bader, G.D. and Ideker, T. (2012) A travel guide to cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.
- Finn, R.D., Mistry, J., Tate, J., Coghill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure-pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
- Janin, J. and Rodier, F. (1995) Protein-protein interaction at crystal contacts. *Proteins*, **23**, 580–587.
- Carugo, O. and Argos, P. (1997) Protein-protein crystal-packing contacts. *Protein Sci.*, **6**, 2247–2263.
- Chakrabarti, P. and Janin, J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.
- Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
- Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
- Zhang, Y. and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
- Charpentier, M., Brouillet, S. and Pothier, J. (2005) YAKUSA: a fast structural database scanning method. *Proteins*, **61**, 137–151.
- Ghoorah, A.W., Devignes, M.D., Smail-Tabbone, M. and Ritchie, D.W. (2013) Protein docking using case-based reasoning. *Proteins*, [doi:10.1002/prot.24433, epub ahead of print].
- Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.D. and Ritchie, D.W. (2010) HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res.*, **38**, W445–W449.
- Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
- Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. and Wilmanns, M. (2000) Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.
- Douangamath, A., Walker, M., Beismann-Driemeyer, S., Vega-Fernandez, M.C., Sterner, R. and Wilmanns, M. (2002) Structural evidence for ammonia tunneling across the $(\beta\alpha)_8$ barrel of imidazole glycerol phosphate synthase bienzyme complex. *Structure*, **10**, 185–193.
- Wierenga, R.K. (2001) The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.*, **492**, 193–198.
- Strohmeier, M., Raschle, T., Mazurkiewicz, J., Rippe, K., Sinning, I., Fitzpatrick, T.B. and Tews, I. (2006) Structure of a bacterial pyridoxal 5'-phosphate synthase complex. *Proc. Natl Acad. Sci. USA*, **103**, 19284–19289.
- Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. and Nussinov, R. (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief. Bioinform.*, **10**, 217–232.
- Kim, W.K., Henschel, A., Winter, C. and Schroeder, M. (2006) The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput. Biol.*, **2**, 1151–1164.
- Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
- Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutillier, K., Burgess, E. *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
- Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernysky, A., Schlessinger, A. *et al.* (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53**(Suppl 6), 430–435.
- Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.