BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 20 2011, pages 2820–2827
doi:10.1093/bioinformatics/btr493

Structural bioinformatics
Advance Access publication August 27, 2011

Spatial clustering of protein binding sites for template based protein docking

Anisah W. Ghoorah[1], Marie-Dominique Devignes[2], Malika Smaïl-Tabbone[3]
and David W. Ritchie[1,*]
[1]INRIA, [2]CNRS and [3]Nancy Université, Orpailleur Team, LORIA, Campus Scientifique, BP 239, 54506
Vandoeuvre-lès-Nancy, France

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: In recent years, much structural information on protein domains and their pair-wise interactions has been made available in public databases. However, it is not yet clear how best to use this information to discover general rules or interaction patterns about structural protein–protein interactions. Improving our ability to detect and exploit structural interaction patterns will help to provide a better 3D picture of the known protein interactome, and will help to guide docking-based predictions of the 3D structures of unsolved protein complexes.

Results: This article presents KBDOCK, a 3D database approach for spatially clustering protein binding sites and for performing template-based (knowledge-based) protein docking. KBDOCK combines residue contact information from the 3DID database with the Pfam protein domain family classification together with coordinate data from the Protein Data Bank. This allows the 3D configurations of all known hetero domain–domain interactions to be superposed and clustered for each Pfam family. We find that most Pfam domain families have up to four hetero binding sites, and over 60% of all domain families have just one hetero binding site. The utility of this approach for template-based docking is demonstrated using 73 complexes from the Protein Docking Benchmark. Overall, up to 45 out of 73 complexes may be modelled by direct homology to existing domain interfaces, and key binding site information is found for 24 of the 28 remaining complexes. These results show that KBDOCK can often provide useful information for predicting the structures of unknown protein complexes.

Availability: http://kbdock.loria.fr/

Contact: Dave.Ritchie@inria.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on June 11, 2011; revised on August 5, 2011; accepted on August 22, 2011

1 INTRODUCTION

Protein–protein interactions (PPIs) are central to many cellular processes. Proteins often perform their function by interacting with other proteins to form protein–protein complexes. In order to understand and predict PPIs reliably, and to relate such interactions to biological function, knowledge of the 3D structures of protein–protein complexes is vitally important. To date, over 65 000 protein structures have been deposited in the Protein Data Bank (PDB; Berman *et al.*, 2002). However, it has been estimated recently that only ∼12% of these structures correspond to heteromeric complexes (Stein *et al.*, 2011). Therefore, to bridge this gap, there is much interest in developing computational techniques to predict how two proteins fit together to form a complex (Aloy *et al.*, 2005). However, until recently, many of the hetero complexes in the PDB have been enzyme-inhibitor complexes, which are relatively easy to model directly. Hence, so-called template-based protein docking has not yet attracted much attention from the research community (Kundrotas *et al.*, 2008).

Since it is well known that protein folds are often more evolutionarily conserved than their sequences (Chothia and Lesk, 1986), and since it has been shown that proteins with similar sequences often interact in similar ways (Aloy *et al.*, 2003), it follows that close structural homologues should also be expected to interact in similar ways. Several studies have found that the locations of protein interaction sites are often conserved, especially within domain families, regardless of the structures of their binding partners (Gunther *et al.*, 2007; Keskin *et al.*, 2005; Korkin *et al.*, 2005, 2006). Additionally, it has also been observed that many protein families employ only one or a small number of binding sites (Keskin and Nussinov, 2007; Shoemaker *et al.*, 2006), suggesting that the same surface patch is often re-used. Indeed, it has been demonstrated previously that the structure of an unknown protein complex may often be successfully modelled using the known binding sites of homologous domains (Kundrotas *et al.*, 2008; Launay and Simonson, 2008). This may be described as template-based docking or docking by homology (Korkin *et al.*, 2006; Kundrotas and Alexov, 2006).

In recent years, much structural information on protein domains and on PPIs has been made available in on-line databases (Tuncbag *et al.*, 2009). However, beyond listing the residues observed at the interface between a given pair of proteins or protein domains, there is no generally accepted way to define what actually constitutes a protein binding site or to quantify whether or not two binding sites are structurally similar. For example, recent methods to compare structural interfaces have used techniques based on e.g. geometric hashing of cliques of interface Cα atoms (Keskin *et al.*, 2004), combining geometric hashing with a physicochemical complementarity scoring function

*To whom correspondence should be addressed.

(Shulman-Peleg *et al.*, 2004), geometric overlap and face angle scores of residue contact vectors (Kim *et al.*, 2006), principal component analysis of residue contact matrices (Aung *et al.*, 2008), complete linkage hierarchical clustering of groups of interface residues (Stein *et al.*, 2010) and dynamic programming-based fragment assembly (Gao and Skolnick, 2010).

Some recent examples of structural PPI databases are PIBASE (Davis and Sali, 2005), SCOPPI (Winter *et al.*, 2006), SCOWLP (Teyra *et al.*, 2006), 3D Complex (Levy *et al.*, 2006), 3D-partner (Chen *et al.*, 2007), PiSite (Higurashi *et al.*, 2009), IBIS (Shoemaker *et al.*, 2010) and 3DID (Stein *et al.*, 2010). Several of these databases describe PPIs in terms of domain–domain interactions (DDIs) because protein domains may often be identified as structural and functional units. Three widely used domain definitions are Pfam (Finn *et al.*, 2010), SCOP (Murzin *et al.*, 1995) and CATH (Cuff *et al.*, 2009). Pfam defines domain using sequence similarities, while the SCOP and CATH domain definitions are based on both sequence and structural similarities. Current structural PPI databases clearly constitute useful bioinformatics resources. However, it is not yet straightforward to use them to extract general patterns that describe the spatial nature of PPIs at the family level. Furthermore, because the question of how best to compare and cluster protein interfaces remains an open problem, it is not yet clear how best to use structural databases to propose suitable templates for homology-based docking predictions.

Here we present KBDOCK, a 3D database approach for spatially clustering protein binding sites and for performing template-based (knowledge-based) protein docking. KBDOCK combines residue contact information from the 3DID database with the Pfam protein domain family classification and protein coordinate data from the PDB in order to superpose and spatially cluster all known hetero DDIs for each Pfam family. The main features that distinguish KBDOCK from existing structural PPI databases are that: (i) it uses the Pfam consensus sequence to guide structural alignments; (ii) it places all of the complexes involving a given Pfam domain family into a common coordinate frame in order to locate the interaction partners consistently in 3D space; (iii) it uses the notion of 'core' and 'rim' interface residues to help define the geometric centre of a binding site; (iv) for each domain of interest, it spatially clusters a weighted combination of the core and rim interface residues of all DDIs involving that domain in order to define domain family binding sites; (v) it may be used to identify automatically the best available DDI template to use to model by homology a complex of two given domains; and (vi) even when no suitable DDI template exists, it can still propose candidate binding sites on one or both interaction partners as potential constraints for computational docking. Thus, KBDOCK represents a novel knowledge-based approach for proposing structural templates for protein docking.

Our approach is illustrated using 10 example query domains, each having multiple hetero interactions that may be clustered into a small number of domain family binding sites. The utility of the approach for template-based protein docking is demonstrated using 73 complexes from the Protein Docking Benchmark (version 4). Overall, up to 45 out of 73 complexes may be modelled by direct homology to existing domain interfaces, and key binding site information is found for 24 of the 28 remaining complexes. There are only four targets for which no homologous hetero DDIs exist. These results show that KBDOCK can often provide useful information for predicting the structures of unknown protein complexes.

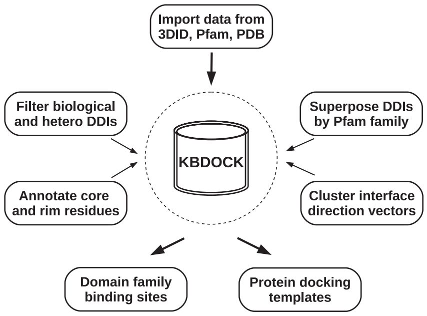
Import data from 3DID, Pfam, PDB
Filter biological and hetero DDIs
Superpose DDIs by Pfam family
KBDOCK
Annotate core and rim residues
Cluster interface direction vectors
Domain family binding sites
Protein docking templates

Fig. 1. Overview of the main KBDOCK data sources and processing steps.

2 METHODS

2.1 Overview of KBDOCK

KBDOCK is built from three main data sources. Multiple sequence alignments and consensus sequences are provided by Pfam, residue contact data and Pfam domain assignments are extracted from the 3DID database, and protein coordinates are obtained from the PDB. We use 3DID as our source of DDIs because it uses the Pfam classification to describe domains, and because it is one of the most complete and up-to-date structural PPI databases currently available. 3DID considers an interface to exist between two domains whenever five or more contacts (hydrogen bonds, electrostatic, or van der Waals atomic interactions) exist between the two domains. This means that 3DID contains both permanent and transient DDIs which may arise from both biologically relevant and non-biological (i.e. crystal contact) interactions. The version of 3DID used here (November 2009) contains a total of 140 612 DDIs drawn from 29 922 PDB structures. A total of 3755 different Pfam families are involved in at least one DDI.

The KBDOCK database is implemented using the MySQL relational database (http://www.mysql.com). All calculations and queries against the database are made using a small set of Prolog programs (http://www.swi-prolog.org/) and R scripts (http://www.r-project.org/). A web interface (http://kbdock.loria.fr) has been implemented using the PHP scripting language (http://php.net) and the Jmol plug-in for visualization (http://jmol.sourceforge.net). Figure 1 summarizes the processing steps used to populate the KBDOCK database. These are described in further detail below. The current version of KBDOCK stores Pfam domain family binding site information for a total of 2721 non-redundant (NR) hetero DDIs involving 1029 Pfam domain families. A MySQL dump of the database is available from the authors on request.

2.2 Selecting non-redundant hetero DDIs

Although the 3DID database stores all known DDIs, our main goal is to predict the 3D structures of heteromeric PPIs, as these are often the most difficult structures to solve experimentally (Ezkurdia *et al.*, 2009). Therefore, for each protein domain present in 3DID, all DDIs involving that domain are extracted and classified as either 'intra', 'homo' or 'hetero'. We consider a DDI to be intra if the interacting domains belong to a single protein chain, and homo if the interacting domains belong to different instances of the same protein chain in a given PDB structure. Otherwise, the interaction is considered to be hetero. Figure 2 illustrates these types of domain interactions schematically. Here, only hetero DDIs are considered further, although in principle the approach could also be used to model homo dimers.

Next, non-biological hetero interactions are filtered out. It has been shown that biological interactions usually have larger interfacial areas than non-biological interactions (Janin and Rodier, 1995). Hence, we use the DSSP program (Kabsch and Sander, 1983) to calculate the solvent accessible surfaces (SASs) buried within each domain interface. If a given domain has multiple interactions with other identical domains, e.g. due to crystal packing,

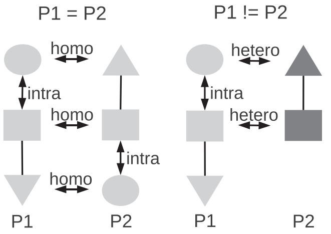

Fig. 2. Schematic illustration of the different types of DDI that may occur between two protein chains, P1 and P2. Protein chains can contain one or more domains connected by linker regions (straight lines). Each shape (circle, rectangle, triangle) represents a different Pfam domain. Lines with arrowheads represent DDIs.

we assume that the interaction with the largest buried SAS corresponds to the biological interaction, and only this DDI is retained.

It is also important to detect and eliminate duplicate or near-duplicate DDIs that may arise in other ways. For example, the same protein complex might have been solved under different crystallographic conditions, or a single crystal structure can sometimes contain different copies of the same complex. In order to deal with such cases, the sequences of the DDI partners are concatenated, and the NRDB90 program (Holm and Sander, 1998) is used with a similarity threshold of 99% to collect a final list of distinct NR DDIs. It is worth noting that because we consider every structure to be useful, a high similarity threshold is used in order to retain as many non-duplicate structures as possible. This does not introduce any bias because here binding sites are defined by spatial clustering and not by counting residue frequencies.

2.3 Annotating DDI interfaces

As discussed above, the residues in PPI sites are often conserved across domain families. Indeed, due to evolutionary pressure, active site residues are often less likely to undergo mutation than other residue positions (Zvelebil *et al.*, 1987), and this phenomenon has been exploited previously to predict molecular interaction sites (Aytuna *et al.*, 2005; Lichtarge *et al.*, 1996). Therefore, the representative sets of hetero DDIs stored in KBDOCK are annotated with both 1D sequence information from Pfam and 3D structure information calculated using DSSP, respectively. For each Pfam domain family, the Pfam database provides a multiple sequence alignment and a consensus sequence of all UniProt sequences belonging to that family. We follow the Pfam convention of considering a residue to be conserved if at least 60% of the amino acids at a given position in the multiple sequence alignment are of the same amino acid type. However, because Pfam uses UniProt sequences rather than PDB structures, and because PDB structures may contain gaps or unresolved regions, we align each PDB sequence with its Pfam/UniProt sequence in order to map every PDB residue to its corresponding Pfam consensus position. This mapping allows the 1D Pfam consensus information to be transferred to each PDB residue position.

In order to enhance the 1D domain family information with 3D interaction information, DSSP is used to calculate the change in solvent accessibility for each interaction residue (as defined by 3DID) between the separate and complexed structures of each domain. Here, we use the notion of ‘core’ and ‘rim’ residues, as defined by Chakrabarti and Janin (2002). An interaction residue is considered to be a core interface residue if it loses at least 75% of its accessible surface area on going from the isolated to the complexed structure. Otherwise, it is considered to be a rim interface residue.

2.4 Defining protein domain family binding sites

Our mapping between the Pfam consensus sequence and PDB residue numbers provides a convenient way to identify the conserved residue positions of all domains stored in KBDOCK. Hence, it is straightforward to retrieve the C_α coordinates of the conserved residue positions in a given domain of interest along with the structures of all of the corresponding DDI partners, and to place these in a common coordinate frame using the ProFit (http://bioinf.org.uk) least-squares fitting program.

Superposing all the DDIs involving a given Pfam domain in this way provides a straightforward way to cluster individual binding sites and to identify automatically distinct PPIs in 3D space. For example, for each superposed DDI, the centre of mass, $\underline{C}$, of each binding site is calculated as a weighted average of the corresponding core (75%) and rim (25%) C_α coordinates. By also calculating the all-atom centre of mass $\underline{D}$ for each domain, an interface direction vector, $\underline{V}$, may then be calculated as

$$\underline{V} = (\underline{C} - \underline{D})/|\underline{C} - \underline{D}|. \quad (1)$$

In order to define domain family binding sites automatically, we cluster the dimensionless interface vectors using Ward’s hierarchical clustering algorithm (Ward, 1963). This is illustrated in Supplementary Figure S1. From visual inspection of several example interfaces, we find that a clustering threshold of 0.4 often gives acceptable clusters.

2.5 Finding docking templates

In order to predict the 3D interaction between a pair of proteins, we need to query the database with two or more domains and to calculate the intersection of the results. This broadly corresponds to calculating a spatial join in a conventional relational database. Although one PPI can involve several DDIs, for simplicity only pair-wise DDIs are considered here. This leads to four possible outcomes, namely that the database is found to contain DDIs involving (i) both query domains together; (ii) both domains individually; (iii) just one domain; or (iv) neither domain.

In the first case, which we call a full homology (FH) DDI, the database DDI would be very likely to provide a good template with which to model the unknown interaction. The two query domains could be docked by homology simply by superposing them onto the FH template. If several such DDIs exist in the database, and if they correspond to different binding sites on the query domain(s), KBDOCK selects for each site the DDI with the highest overall sequence identity to the query domains. On the other hand, if homologous DDIs exist in the database for both of the query domains individually (case ii), it is reasonable to suppose that their binding sites might be re-used in the target complex, thus providing a rational way to initialize a more exhaustive computational docking calculation. Similarly, if just one of the target domains has known binding sites (case iii), these could still be used to constrain a computational docking run. These two cases may be termed docking by ‘semi-homology’ (SH) in analogy to the notion of a semi-join in relational algebra. In such cases, KBDOCK selects the best available homologous DDI for one or both query domains, as appropriate, and it identifies the residue(s) on the query domain(s) which lie closest to the centre of the corresponding binding site(s). These residue identities could then be used to define computational docking constraints. Clearly, if the database contains no homologous interactions, the target complex must be modelled by *ab initio* docking. However, as this study is primarily concerned with exploring a new knowledge-based approach for finding docking templates, the use of computational docking techniques is not considered here.

2.6 The Protein Docking Benchmark

In order to explore the utility of using KBDOCK to find homology templates for protein docking, our approach was used to predict a subset of the protein docking targets in version 4 of the Protein Docking Benchmark (Hwang *et al.*, 2010). The Docking Benchmark is a non-redundant expert-curated set of 176 protein complexes for which the bound complex structures, and most of the unbound component structures, have been solved by X-ray crystallography to a resolution of 3.25 Å or better. Since KBDOCK works at the domain level, we selected all single domain complexes belonging to the ‘Enzyme-Inhibitor’ (here called ‘Enzyme’) and ‘Other’ categories of the Docking Benchmark for this preliminary experiment. In other words, for simplicity we exclude the Benchmark ‘Antibody’ complexes (because apart from involving the

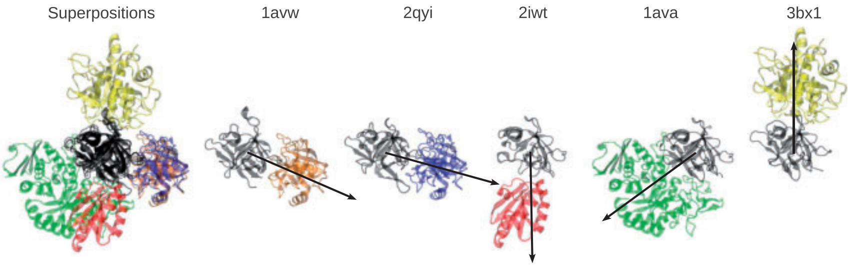

Fig. 3. This figure shows the superpositions and interface direction vectors [Equation (1)] of the five DDIs of the *Kunitz legume* Pfam family. Here, the 1avw (porcine trypsin/soybean trypsin inhibitor) and 2qyi (bovine trypsin/trypsin inhibitor) complexes share a common binding site, and clearly have very similar interface vectors.

antibody hypervariable loops, antibody–antigen interactions generally do not entail homology) and we exclude all other complexes involving multiple domains. This gives a test set of 36 Enzyme and 37 Other target complexes.

It should be noted that the Docking Benchmark complexes do not necessarily provide an unbiased set of homology modelling targets. Because several of the benchmark proteins have been relatively well studied, it is possible that the PDB could contain more homologues of those complexes than randomly selected complexes. In order to take into account this possible source of bias, a stringent test would be to exclude as templates all structures with more recent PDB deposition dates than the target structure. However, filtering complexes by target date often excludes a large proportion of the database. Therefore, in order to provide upper and lower bounds on the utility of template-based modelling, and to try to quantify the growing usefulness of knowledge-based approaches, we report results both with and without date filtering.

3 RESULTS

3.1 Defining domain family binding sites

Superposing families of related DDIs in a common coordinate frame and clustering their interface direction vectors [Equation (1)] provide a straightforward way to analyse structural relationships between the members of a given query domain. Figure 3 shows the superpositions and interface vectors calculated for the five DDIs involving the *Kunitz legume* Pfam family. This figure clearly shows that this domain has four distinct interaction sites, one of which is common to two different trypsin/inhibitor complexes. Supplementary Figure S1 shows the spatial clustering dendrogram for this family, and for a further three example Pfam families (namely, *Kunitz BPTI*, *Ribonuclease* and *Actin*).

Spatial clusters have been calculated and stored in KBDOCK for all the 1029 Pfam domain families which are involved in hetero interactions. Superposing and clustering all Pfam domain binding sites in KBDOCK takes ~8 CPU hours on a 64-bit 2.8 GHz Q9550 processor. Table 1 summarizes the number of hetero DDI partners and calculated binding sites for 10 example Pfam domain families, including the four examples considered above. This table shows that these Pfam domains typically have from one to four binding sites, according to our spatial clustering algorithm. It is interesting to note that even domains involved in many DDIs such as *Kunitz BPTI*, *Trypsin* and *Actin* still have only a relatively small number of distinct binding sites.

Figure 4 shows the DDI superpositions for the 10 Pfam families listed in Table 1. In most cases, visual inspection of the complexes in this figure readily confirms the calculated number of binding sites given in Table 1. For example, the *Potato inhibit* domain interacts with eight other domains (all serine proteases) using a single binding site. On the other hand, the *Kunitz BPTI* domain has two inhibitory binding sites, and, as shown in Table 1, the *Kunitz legume* inhibitor has four binding sites that form distinct interfaces with four different domain families, namely *Trypsin*, *Thioredoxin*, *Alpha-amylase* and *Peptidase S8*. Conversely, *Thioredoxin* interacts with eight different Pfam families, but it does so using just two overlapping binding sites.

Table 1. Summary of the number of DDIs and calculated binding sites for 10 example Pfam domains stored in KBDOCK

Pfam ID	Pfam name	Function	No. of DDIs	No. of binding sites
PF00197	Kunitz legume	Protease inhibitor	5	4
PF00014	Kunitz BPTI	Protease inhibitor	27	2
PF00280	Potato inhibit	Protease inhibitor	8	1
PF00089	Trypsin	Protease	98	6
PF00062	Lys	Hydrolase	10	5
PF00545	Ribonuclease	Hydrolase	9	1
PF00022	Actin	Protein binding	24	4
PF00059	Lectin C	Glycoprotein binding	14	4
PF00111	Fer2	Ferredoxin	14	3
PF00085	Thioredoxin	Redox protein	8	2

For domains that have multiple binding sites and which interact with several different domain partners (e.g. *Fer2*, *Lectin C*, *Lys*, *Actin* and *Trypsin*), it can be difficult to distinguish all the interactions visually. Hence, KBDOCK allows the user to select and display only those DDIs involving a given binding site. Comparing the DDIs of binding sites selected in this way using 3D graphical visualization software such as Jmol often shows that our clustering algorithm calculates acceptable clusters in almost all cases.

Figure 5 shows the distribution and the change with time of the number of binding sites per domain family (excluding the very large *C1-set* immunoglobulin domain family) of all NR hetero DDIs in KBDOCK. This figure confirms that most domains typically have from one to four hetero binding sites, and only a very small number of domains such as *Trypsin* (six binding sites) have more than this. Indeed, over 60% of all hetero domains in KBDOCK have just one binding site, which supports the notion that domain binding sites are often re-used in different DDIs. It is interesting to note that despite the growing number of Pfam domains for which KBDOCK contains

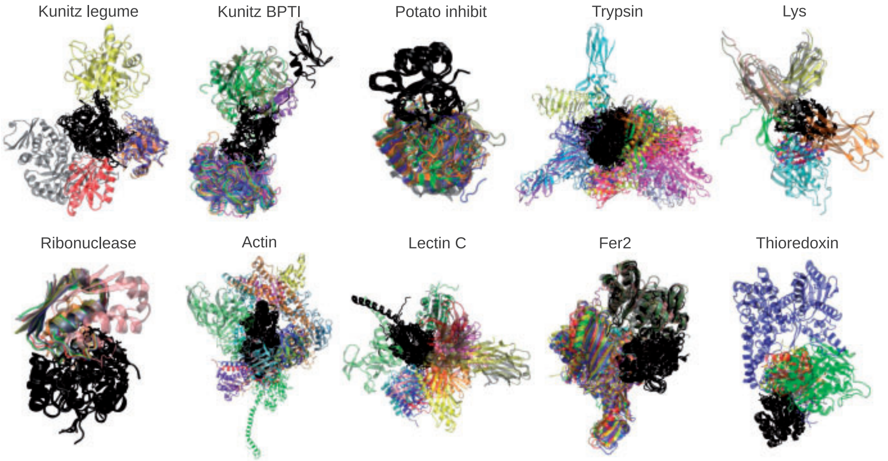

Fig. 4. DDI superpositions for 10 example Pfam domains (Table 1) in the coordinate frame of the query. In each case, the query domain is shown in black.

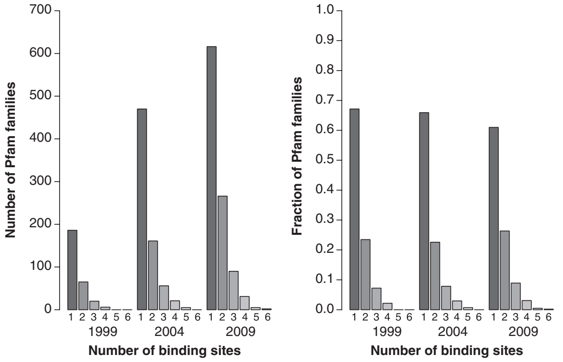

Fig. 5. The calculated number of hetero binding sites per domain family by PDB deposition date for all Pfam families except the *C1-set* immunoglobulin domains.

hetero complexes, the relative proportion of domains having 1, 2, 3, or 4 binding sites seems remarkably stable.

Table 2. Summary of the KBDOCK template modelling results for the 73 selected Protein Docking Benchmark targets[a]

Target class	Total targets	FH templates	Two SH templates	One SH template	No templates
No date filtering					
Enzyme	36	24/24	(3 + 1)/5	3/5	2
Other	37	21/21	(0 + 0)/3	5/11	2
Total	73	45/45	(3 + 1)/8	8/16	4
With date filtering					
Enzyme	36	13/13	(2 + 1)/5	7/11	7
Other	37	13/13	(0 + 0)/1	8/15	8
Total	73	26/26	(2 + 1)/6	15/26	15

[a] This table shows the number of docking targets for which the proposed templates are correct compared to the total number of templates retrieved. When SH templates are found for both domains individually ('Two SH templates'), the figures in brackets give the number of cases in which both binding sites are modelled correctly *plus* the number in which only one binding site is modelled correctly. Full details are presented in Supplementary Tables S1–S3.

3.2 Docking by homology

Table 2 summarizes the results of querying KBDOCK to find docking templates for the 73 target complexes, both with and without filtering DDIs by PDB deposition date. Full details of these results are given in Supplementary Tables S1–S3. Supplementary Figure S2 shows two examples of docking targets for which KBDOCK finds FH and SH templates, respectively. For all targets, the structures of the unbound domains given by Hwang *et al.* (2010) were used as query domains, and the corresponding crystallographic complex (i.e. the expected solution) was excluded from the modelling procedure. Here, a FH template is considered to be correct if the root mean squared deviation (RMSD) between it and the native complex is <10 Å. This is similar to the CAPRI criteria for an 'acceptable' docking prediction (Mendez *et al.*, 2005). According to this criterion, Table 2 shows that KBDOCK finds good FH templates for a total of 24 out of 36 Enzyme target complexes, although this number falls to 13 when PDB deposition date filtering is applied. A further 10 targets have SH DDIs involving one or both of the target domains, and just two Enzyme targets have no hetero DDI information. Similarly, Table 2 shows that 21 (or 13 with date filtering) of the 37 Other targets may be modelled using FH templates, and a further 14 targets have SH DDIs involving one or both of the target domains. Like the Enzyme targets, only two of the Other targets have no hetero DDI information.

Table 2 shows that all the retrieved FH templates are correct according to the 10 Å RMSD threshold. Thus, if KBDOCK retrieves a FH template, there is a high probability that it represents a good model of the target complex. However, Supplementary Table S1 shows that KBDOCK sometimes finds more than one distinct FH interface for a given pair of query domains. In other words, the instances of two domain families can sometimes interact via more than one combination of binding sites. For example, KBDOCK retrieves two FH templates for three of the Enzyme targets (1dfj, 1eaw, 2pcc) without date filtering, and for just one target (1eaw) when date filtering is applied. Subsequent visual inspection of the calculated templates for the matriptase/BPTI target (PDB code 1eaw) showed that the first trypsinogen/BPTI DDI (2r9p) provides a very good template (with an overall RMSD between the template and target of 0.79 Å), whereas the second (lower sequence identity) FH template corresponds to a different inhibitor orientation found in the prothrombin/boophilin complex (2ody; 8.54 Å RMSD).

Similarly, visual inspection of the calculated templates for two of the Other targets (1fqj, 1ml0) confirmed that these FH templates correspond to two different binding modes for both the G-protein complex (1fqj) and the M3-protein complex (1ml0), and that the first (highest sequence identity) template best matches the target (0.70 and 1.03 Å RMSD, respectively). On the other hand, the two DDIs (1z7x and 2bex) calculated for the large *RnaseA/LRR 1* Enzyme complex (1dfj) were seen to overlap considerably, and the two large binding sites calculated for the *RnaseA* domain should have been clustered as a single binding site. Similarly, two of the Other DDIs (1mq8, 2ayo) are calculated to have two distinct binding sites, although visual inspection again suggests that these should have been clustered as a single binding site. We believe that such clustering artefacts sometimes arise due to different assignments of core and rim residues in different instances of homologous DDIs. The peroxidase/cytochrome C Enzyme target (2pcc) is another interesting case. Although the two DDIs calculated for this target are quite distinct, further investigation revealed that one of the DDIs arises, because the crystal contact between these domains was larger than the biological contact in the 1s6v structure. Consequently, two binding sites instead of one were also calculated for these domains. Thus, KBDOCK can successfully retrieve alternate FH binding modes when they exist in the database, but it can also be seen that its clustering algorithm has a slight tendency to overestimate the number of distinct binding sites.

As might be expected, fewer FH templates are available when PDB data filtering is applied, and this causes an increase in the number of proposed SH templates. When only SH templates are retrieved, we assess their quality by comparing each proposed binding site with that of the native complex and if our interface clustering algorithm would group them together, we consider the retrieved template to be correct. The final two columns of Supplementary Tables S2 and S3 show the outcome of this test, and Table 2 summarizes the overall results. For example, SH DDIs involving the two individual query domains exist for five of the Enzyme targets. Supplementary Table S2 shows that three of these targets (1e6e, 1acb, 1f6m) may be modelled correctly by re-using their Pfam domain binding sites, and one further target (1ov8) may be partially modelled by re-using one of the two proposed SH templates. On the other hand, there are three Other targets for which the two query domains both have SH templates, but none of these lead to acceptable models. For those cases where only one SH template exists for a given target, the binding sites of one of the queries is found to be re-used in the target DDI in a total of three out of five Enzyme targets (1gl1, 4cpa, 1fq1) and five out of 11 Other targets (1ktz, 2g77, 1wq1, 2h7v, 1y64). In order to assist any subsequent computational docking calculation that might use these templates, Supplementary Tables S2 and S3 show the proposed PDB template along with the name of the query residue calculated to be at the centre of the binding site.

Overall, it can be seen that KBDOCK can provide high-quality FH docking templates for a total of 45 of the 73 targets (or 26/73 with date filtering). Even when no FH templates exist, KBDOCK can still find useful binding site information for at least one of the domain partners for 12 of the remaining 28 targets (or 18/47 with date filtering). These results demonstrate that the approach embodied in KBDOCK provides a useful way to find protein docking templates.

4 DISCUSSION

4.1 Comparison with previous approaches

Because the main aim of KBDOCK is to facilitate automatic docking by homology, it has several novel aspects that have not been explored in previous studies of structural PPIs. In particular, because protein docking is inherently a spatial problem (with six degrees of freedom in the simplest rigid body assumption), KBDOCK was designed from the start to consider the relative spatial arrangements of interacting protein domains, and to deal with cases where a full homology template is not necessarily available. This is in contrast to the most previous PPI classification approaches, which generally apply clustering techniques to groups of residues belonging to both partners of existing interfaces. For example, 3DID defines a domain interface by applying complete linkage hierarchical clustering to identify groups of shared interface residues within a Pfam domain family (Stein *et al.*, 2009). It then labels each distinct occurrence of a domain/interface pair as an ‘interaction topology’, and it applies a further round of hierarchical clustering to define ‘global interface clusters’ that group together individual interfaces (Stein *et al.*, 2010). This gives an average of about 10 global interfaces per Pfam domain [see Figure 3 of Stein *et al.* (2010)].

Kim *et al.* (2006) represent an interface as a pair of ‘face vectors’, each of which contains a list of ones and zeros to represent the contacting and non-contacting residues of each domain, respectively. The face vectors within a SCOP domain are then grouped according to the cosine similarity between their face vectors, and interfaces with similar faces are superposed and clustered according to their face overlap and the angle between the centroids of pairs of faces (Kim *et al.*, 2006; Winter *et al.*, 2006). Pairs of faces are then combined to define interface types. This procedure is reported to give on average about 5.4 distinct interface types per SCOP domain family (Winter *et al.*, 2006). Other approaches such as 3DID, PPiClust (Aung *et al.*, 2008) and I2I-SiteEngine (Shulman-Peleg *et al.*, 2004) also cluster and analyse pair-wise interfaces rather than individual binding sites. Clearly, the interface direction vectors used in KBDOCK share a similar inspiration to the face angle measure of Kim *et al.* (2006). However, Kim *et al.* (2006) focused on studying the diversity of domain interfaces, the evolution of hub proteins and gene fusion events (often manifested as intra-domain interactions), whereas our study focuses on finding docking templates for hetero domain interactions. Thus, our study complements and extends previous work.

Previous template-based docking approaches have used comparative patch analysis, threading and sequence alignment techniques (Chen and Skolnick, 2008; Korkin *et al.*, 2006; Kundrotas and Vakser, 2010; Kundrotas *et al.*, 2008; Launay and Simonson, 2008), for example. Hence, at a conceptual level, KBDOCK shares a similar inspiration with the comparative patch analysis approach of Korkin *et al.* (2006). This approach defines and clusters binding sites of interacting SCOP domains using a scalar ‘localisation index’ calculated as a sum of contact residue frequencies in the context of the superposed domains of a given

SCOP family (Korkin *et al.*, 2005). This index serves as a kind of fuzzy set membership measure, and does not consider the directional nature of the interface, whereas KBDOCK explicitly clusters binding sites according to the spatial orientation of their core and interface residues.

From a template docking point of view, KBDOCK is somewhat similar to the HOMBACOP approach of Kundrotas *et al.* (2008). HOMBACOP begins by using PSI-BLAST to identify candidate structural templates for a given pair of sequences, and these are refined using a further round of sequence-based template matching using a position-specific scoring matrix enriched with interface information of the known templates. In a similar spirit, Launay and Simonson (2008) use the solvent accessibility of interface residues to enhance their Needleman–Wunsch alignment of candidate templates. However, they then use an energy function to select the final template, whereas HOMBACOP uses sequence similarity and KBDOCK uses structural similarity to the target domains as the final selection criteria. Compared with HOMBACOP that used the PROTCOM database (Kundrotas and Alexov, 2007) without date filtering to produce 19 models for 43 targets (44%) from the Docking Benchmark version 2 (Mintseris *et al.*, 2005), KBDOCK finds good FH templates for 26 (36%) and 45 (62%) out of 73 targets with and without date filtering, respectively. Hence, KBDOCK appears to be rather competitive compared to the earlier approach.

Overall, KBDOCK provides high-quality FH docking templates for 62% of the targets studied here, and it finds useful binding site information for a further 39% (11/28) of the remaining targets. Following these very promising results, we are extending KBDOCK to deal with multi-domain complexes, and to link it directly to our rigid body docking software (Ritchie and Kemp, 2000).

4.2 Implications for the 3D interactome

There is growing interest in using docking techniques to predict large-scale structural PPIs (Kundrotas *et al.*, 2010; Launay and Simonson, 2008; Mosca *et al.*, 2009; Sinha *et al.*, 2008; Wass *et al.*, 2011). However, results from the CAPRI docking experiment (Lensink and Wodak, 2010) show that current docking algorithms still face the problem of how to distinguish a good solution from a list of feasible but mostly incorrect predicted docking orientations. On the other hand, exploiting biochemical or biophysical knowledge in data-driven docking (van Dijk *et al.*, 2005) can often help to constrain the scope of a docking calculation and considerably improve the quality of the results (Korkin *et al.*, 2006; Lensink and Wodak, 2010; Ritchie, 2008). Hence, if prior biological knowledge is available in a suitable form, it would be desirable to be able to incorporate it automatically in a docking calculation.

Kim *et al.* (2006) note that many of the currently known interface types only started to become available in the mid-1990s. Hence, early docking and interface studies only had a small repertoire of interface types to work with. They also found that although the number of interface types continues to grow, the rate of growth is currently much less than the growth in the total number of multi-domain structures that are being solved (Figure 5 of Kim *et al.*, 2006). Our analysis of the rate of growth in the number of hetero binding sites since 1999 (Fig. 5) also shows only a modest increase in the number of Pfam families having multiple hetero binding sites, despite over a 3-fold increase in the number of Pfam families for which hetero complexes are now available. This strongly supports the notion that protein binding sites are very often re-used. Of course, the hetero complexes available in the PDB are not necessarily representative of the whole structural interactome. Nonetheless, if the very small numbers of hetero protein binding sites found here do indeed turn out to be typical, this will have considerable implications for future data-driven and template-based docking approaches, and for populating 3D PPI networks on a genomic scale.

5 CONCLUSION

KBDOCK provides a systematic way to store and analyse the 3D structures of protein domain binding sites. By superposing the structures of all hetero DDIs involving a given query domain, and by using the simple notion of an interface direction vector to define the central region a protein binding site, a small number of spatially distinct binding sites may be identified for each Pfam domain family. Using this approach, we find that the majority of the 1029 Pfam domain families have a small number (up to four) of hetero binding sites, and over 60% have just one hetero binding site.

KBDOCK can be used to find automatically homologous hetero DDIs with which to model the unknown 3D structure of given protein complex. In 60% of the docking benchmark examples studied, KBDOCK finds a small number of high quality DDI templates with which to model the target complex. Furthermore, one of the unique strengths of KBDOCK is that it can find semi-homologous templates even when no full homology template is available. Hence, KBDOCK provides a useful knowledge-based approach for template-based protein docking and for helping to describe and understand structural PPIs on a genomic scale.

ACKNOWLEDGEMENT

We thank Matthieu Chavent for useful discussions.

Funding: Agence Nationale de la Recherche, grant reference ANR-08-CEXC-017-01.

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.

Aloy,P. *et al.* (2005) Protein complexes: structure prediction challenges for the 21st century. *Curr. Opin. Struct. Biol.*, **15**, 15–22.

Aung,Z. *et al.* (2008) PPiClust: efficient clustering of 3D protein-protein interaction interfaces. *J. Bioinformatics Comput. Biol.*, **6**, 415–433.

Aytuna,A.S. *et al.* (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, **21**, 2850–2855.

Berman,H.M. *et al.* (2002) The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **58**, 899–907.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins Struct. Funct. Genet.*, **47**, 334–343.

Chen,H. and Skolnick,J. (2008) M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys. J.*, **94**, 918–928.

Chen,Y.C. *et al.* (2007) 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res.*, **35**, W561–W567.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

Cuff,A.L. *et al.* (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.

Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.

Ezkurdia,L. *et al.* (2009) Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinformatics*, **10**, 233–246.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Gao,M. and Skolnick,J. (2010) iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics*, **26**, 2259–2265.

Gunther,S. *et al.* (2007) Docking without docking: ISEARCH – prediction of interactions using known interfaces. *Proteins Struct. Funct. Bioinformatics*, **69**, 839–844.

Higurashi,M. *et al.* (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, **37**, D360–D364.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Hwang,H. *et al.* (2010) Protein-protein docking benchmark version 4.0. *Proteins Struct. Funct. Bioinformatics*, **78**, 3111–3114.

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

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Keskin,O. and Nussinov,R. (2007) Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, **15**, 341–354.

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

Keskin,O. *et al.* (2005) Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, **345**, 1281–1294.

Kim,W.K. *et al.* (2006) The many faces of protein-protein interactions: a compendium of interface geometry. *PLoS Comput. Biol.*, **2**, 1151–1164.

Korkin,D. *et al.* (2005) Localization of protein-binding sites within families of proteins. *Protein Sci.*, **14**, 2350–2360.

Korkin,D. *et al.* (2006) Structural modeling of protein interactions by analogy: application to PSD-95. *PLoS Comput. Biol.*, **2**, e153.

Kundrotas,P.J. and Alexov,E. (2006) Predicting 3D structures of transient protein-protein complexes by homology. *BBA Proteins Proteomics*, **1764**, 1498–1511.

Kundrotas,P.J. and Alexov,E. (2007) PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res.*, **35**, D575–D579.

Kundrotas,P.J. and Vakser,I.A. (2010) Accuracy of protein-protein binding sites in high-throughput template-based modeling. *PLoS Comput. Biol.*, **6**, e1000727.

Kundrotas,P.J. *et al.* (2008) Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. *Int. J. Biol. Macromol.*, **43**, 198–208.

Kundrotas,P.J. *et al.* (2010) GWIDD: genome-wide protein docking database. *Nucleic Acids Res.*, **38**, D513–D517.

Launay,G. and Simonson,T. (2008) Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics*, **9**, 427.

Lensink,M.F. and Wodak,S.J. (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinformatics*, **78**, 3073–3084.

Levy,E.D. *et al.* (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, 1395–1406.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Mendez,R. *et al.* (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins Struct. Funct. Bioinformatics*, **60**, 150–169.

Mintseris,J. *et al.* (2005) Protein-protein docking benchmark 2.0: An update. *Proteins Struct. Funct. Bioinformatics*, **60**, 214–216.

Mosca,R. *et al.* (2009) Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput. Biol.*, **5**, e1000490.

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.

Ritchie,D.W. (2008) Recent progress and future directions in protein-protein docking. *Curr. Protein Peptide Sci.*, **9**, 1–15.

Ritchie,D.W. and Kemp,G.J.L. (2000) Protein docking using spherical polar Fourier correlations. *Proteins Struct. Funct. Genet.*, **39**, 178–194.

Shoemaker,B.A. *et al.* (2006) Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci.*, **15**, 352–361.

Shoemaker,B.A. *et al.* (2010) Inferred biomolecular interaction server-a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.*, **38**, D518–D524.

Shulman-Peleg,A. *et al.* (2004) Protein-protein interfaces: Recognition of similar spatial and chemical organizations. *Proc. Algorithms Bioinformatics*, **3240**, 194–205.

Sinha,R. *et al.* (2008). Docking by structural similarity at protein-protein interfaces. *Proteins Struct. Funct. Bioinformatics*, **78**, 3235–3241.

Stein,A. *et al.* (2009) 3did update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.

Stein,A. *et al.* (2010) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.

Stein,A. *et al.* (2011) Three-dimensional modeling of protein interactions and complexes is going ’omics. *Curr. Opin. Struct. Biol.*, **21**, 200–208.

Teyra,J. *et al.* (2006) SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **7**, 104.

Tuncbag,N. *et al.* (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief. Bioinformatics*, **10**, 217–232.

van Dijk,A.D. *et al.* (2005) Data-driven docking for the study of biomolecular complexes. *FEBS J.*, **272**, 293–312.

Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.

Wass,M.N. *et al.* (2011) Towards the prediction of protein interaction partners using physical docking. *Mol. Syst. Biol.*, **7**, 469.

Winter,C. *et al.* (2006) SCOPPI: a structural classification of protein–protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.

Zvelebil,M.J. *et al.* (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957–961.