BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 9 2005, pages 1901–1907
doi:10.1093/bioinformatics/bti277

Structural bioinformatics

PIBASE: a comprehensive database of structurally defined protein interfaces

Fred P. Davis[1,2,3] and Andrej Sali[2,3,*]

[1]Graduate Group in Biophysics, [2]Department of Biopharmaceutical Sciences and [3]Department of Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA 94143, USA

Received on November 19, 2004; revised on January 7, 2005; accepted on January 13, 2005
Advance Access publication January 18, 2005

ABSTRACT

Motivation: In recent years, the Protein Data Bank (PDB) has experienced rapid growth. To maximize the utility of the high resolution protein–protein interaction data stored in the PDB, we have developed PIBASE, a comprehensive relational database of structurally defined interfaces between pairs of protein domains. It is composed of binary interfaces extracted from structures in the PDB and the Probable Quaternary Structure server using domain assignments from the Structural Classification of Proteins and CATH fold classification systems.

Results: PIBASE currently contains 158 915 interacting domain pairs between 105 061 domains from 2125 SCOP families. A diverse set of geometric, physiochemical and topologic properties are calculated for each complex, its domains, interfaces and binding sites. A subset of the interface properties are used to remove interface redundancy within PDB entries, resulting in 20 912 distinct domain–domain interfaces. The complexes are grouped into 989 topological classes based on their patterns of domain–domain contacts. The binary interfaces and their corresponding binding sites are categorized into 18 755 and 30 975 topological classes, respectively, based on the topology of secondary structure elements. The utility of the database is illustrated by outlining several current applications.

Availability: The database is accessible via the world wide web at http://salilab.org/pibase

Contact: sali@salilab.org

Supplementary information: http://salilab.org/pibase/suppinfo.html

INTRODUCTION

Proteins do not act in isolation, but rather through interactions with molecules in their spatio-temporal environment that includes small molecules and nucleic acids, as well as other proteins (Alberts, 1998). Therefore, the structures of individual proteins are often uninformative of biological function if taken out of context. Recent experimental advances have addressed this problem by enabling studies of protein interactions along two frontiers (Sali *et al.*, 2003; Russell *et al.*, 2004): (1) large-scale detection of protein–protein interactions (Fields and Song, 1989; Uetz *et al.*, 2000; Ito *et al.*, 2001; Gavin *et al.*, 2002; Ho *et al.*, 2002) and (2) structure determination of protein complexes (Sali, 2003). To maximize their utility, these experiments require informatics resources to store, organize, visualize, analyze and disseminate the data. The objective is to understand the evolution and physics of protein interactions and to develop predictive models of protein structure and function.

Experimentally determined structures of protein complexes are deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000). The PDB is growing rapidly, in part due to the recent structural genomics effort (Berman *et al.*, 2000; Editorial, 2004). The PDB currently holds ∼28 000 structures. Each entry contains on average 2.2 protein chains, and each chain contains on average 2.1 domains. Domains are considered the basal unit of protein structure, function and evolution (Ponting and Russell, 2002). These units fold independently, often mediate a specific biological function, and combine modularly to form larger proteins. Several approaches to the definition of domain boundaries in proteins have been developed based on sequence and structure (Veretnik *et al.*, 2004). The Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997) are two commonly used structure-based domain definition and classification systems.

Biologically relevant quaternary states are proposed for crystallographic protein structures by the Probable Quaternary Structure (PQS) server (Henrick and Thornton, 1998). The server applies crystallographic and non-crystallographic symmetry operations to the PDB structure, and then assesses the validity of each chain interface using a set of empirically derived cutoffs for properties such as buried solvent accessible surface area, buried number of residues, hydrogen bonding and salt bridges. The PDB and the PQS are sources of the highest resolution protein–protein interaction data.

The structures of protein subunit interfaces have long been studied using collections of protein chain and domain interfaces (Argos, 1988; Janin *et al.*, 1988; Tsai *et al.*, 1996; Jones and Thornton, 1996; Conte *et al.*, 1999; Hitz and Honig, 1999, http://trantor.bioc.columbia.edu/cgi-bin/SPIN/; Park *et al.*, 2001; Aloy *et al.*, 2003; Keskin *et al.*, 2004). Numerous analyses have used datasets of protein chain interfaces extracted from the PDB to investigate properties such as residue type propensities, sequence conservation and structure conservation at protein interfaces (Valdar and Thornton, 2001; Ofran and Rost, 2003; Caffrey *et al.*, 2004; Aloy *et al.*, 2003; Jones *et al.*, 2000; Jones and Thornton, 1996). These studies of interface properties have given valuable insights into the physics and evolution of protein interactions.

In this paper, we describe a comprehensive relational database of structurally defined domain–domain interfaces. We annotate them by

*To whom correspondence should be addressed.

a diverse set of geometric, physicochemical and topologic properties that characterize the structure of the protein complexes from the level of the complex to the atomic level details of each interface. A subset of these properties are used to remove interface redundancy as well as categorize the complexes, the interfaces and the binding sites into topological classes. This multilevel characterization allows queries that span a range from properties of specific interfaces to proteome level views of interactions. The motivation in developing PIBASE has been to create a comprehensive repository of information characterizing the structure of protein complexes at a range of size scales using a diverse set of descriptors.

The construction of the database is described first in the Methods and Results section, detailing the data sources, interface definition, properties computed, interface redundancy removal and clustering (Methods). We then discuss the composition of the database, describing the contents as well as the distributions of several of the computed interface properties (Results). Finally, we conclude with a brief discussion of several of our current applications of PIBASE (Discussion section).

METHODS AND RESULTS

Sources of protein structures and their classification

Two types of input data were used: protein structures and domain definitions. The structures were obtained from the PDB (Berman *et al*., 2000) and the PQS server (Henrick and Thornton, 1998). For those structures determined by NMR spectroscopy, the first model in the ensemble was used.

The domain definitions for the PDB structures were obtained from the SCOP (Murzin *et al*., 1995) and CATH (Orengo *et al*., 1997) classification systems. A mapping was generated between the PDB and PQS chains that allowed domain definitions to be transferred from the PDB to its associated PQS entries. Approximately 1.5% of PQS entries contain chains with sequence changes, chain breaks or chain mergers relative to their parent PDB structure. These differences occur for reasons such as missing density in the PDB structure. Domain definitions were not generated for these PQS entries as the chain mapping is difficult and inexact.

Detection of domain–domain interfaces

The list of binary interfaces was generated by a three-step procedure (Fig. 1). (1) Interatomic distances were calculated for all structures using a user specified distance cutoff. A cutoff of 6.05 Å was chosen unless specified otherwise, to allow contacts made via water molecules (Robert and Janin, 1998; Robert and Ho, 1995). (2) The interatomic distances were then combined with the domain definitions to create a list of all domain pairs that share at least one interatomic distance below the specified distance threshold. This list of interacting domain pairs serves as the core of PIBASE. (3) Buried solvent accessible surface area (below) was also computed for each interacting domain pair and a minimum cutoff on the burial was imposed to yield the list of interfaces. Unless specified otherwise, a cutoff of 300 $Å^2$ was used, as justified below.

It is often difficult to ascertain the biologically relevant quaternary state solely based on crystallographic information (Carugo and Argos, 1997). Previous studies have attempted to determine the biological relevance of observed chain contacts by two alternative strategies. The first is to define empirical thresholds on a set of interface properties (such as change in solvent accessible surface area, number of buried residues, etc.) that are able to distinguish true biological interfaces from crystal packing artifacts (Henrick and Thornton, 1998). The second strategy is to analyze the conservation of residues at the observed interface, with the hypothesis that a biological interface would be more conserved than a crystal packing artifact (Elcock and McCammon, 2001; Valdar and Thornton, 2001). While validation of both strategies is difficult, error rates of 4.3–20% have been reported (Elcock and McCammon, 2001; Henrick and Thornton, 1998; Valdar and Thornton, 2001). The first strategy of empirical cutoffs has been implemented in an automated fashion in the PQS server (Henrick and Thornton, 1998). PQS uses a threshold of 400 $Å^2$ as one of the factors in determining biological relevance of a chain–chain interface. Unless otherwise mentioned, we used a lower threshold of 300 $Å^2$, which removed ~45% of the interacting SCOP domain pairs (Supplementary Figure 1). However, all interacting domain pairs are stored and all analyses can be performed easily with any choice of cutoff.

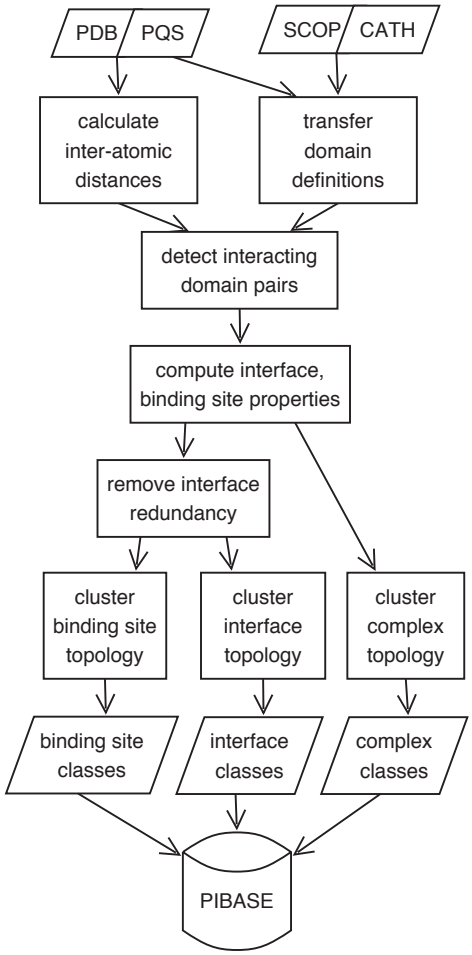

Fig. 1. PIBASE build procedure. Briefly, interatomic distances calculated from PDB and PQS structures are combined with domain definitions from SCOP and CATH to generate a list of interacting domain pairs. A set of properties are then computed for all the interfaces and binding sites. A subset of these properties are then used to remove interface redundancy within structures associated with each PDB code. The binding sites, interfaces and complexes are then clustered using their topological properties.

Properties of complexes, domains, interfaces and binding sites

For each processed structure, PIBASE contains a set of properties at four different levels: the complex, its constituent domains, binary interfaces and binding sites (the two halves of an interface) (Supplementary Table 1). These properties are described next.

Complexes. A domain connectivity graph was generated for each structure, describing the pattern of domain–domain contacts. In this graph, a domain is a node, and a binary domain interface is an edge (Supplementary Figure 2b). This graph captures the arrangement of the individual domains in the complex. A crude linear representation of this graph is then computed and used as a topological fingerprint to group the structures into topological classes (Supplementary Information: Topological Fingerprints, Supplementary Figure 2d). While degeneracy exists in this representation (i.e. two distinct topologies may have the same sorted edge list), it is useful as both a query and crude clustering term.

Domains. Solvent accessible surface area was computed for each domain using a probe radius of 1.4 Å with the algorithm of Richmond and Richards as implemented in MODELLER (Richmond and Richards, 1978; Sali and Overington, 1994). Secondary structure assignments are made by DSSP (Kabsch and Sander, 1983). Classification codes (e.g. class, fold, superfamily and family) from the domain assignment system used, SCOP or CATH, are also associated with each domain.

Interfaces and Binding Sites. The interface and binding site properties compose the majority of PIBASE content. A subset of these properties were used for redundancy removal and clustering of the interfaces. The interface properties can be grouped into two categories: non-contact and contact. Non-contact properties (properties 1–8 in the Interface column of Supplementary Table 1c) are properties that are a sum or a union of the properties in the corresponding domains. For instance, the number of residues presented at the interface is computed as a sum of the number of residues presented by each of the two binding sites. The contact properties (properties 9–17 in the Interface column of Supplementary Table 1c) implicitly capture the interface orientation. These properties cannot be defined independently by the two binding sites. The binding site properties fall into two categories: non-topology and topology. Non-topology properties (properties 1–8 in the Binding site column of Supplementary Table 1c) describe the size and physicochemical properties of the binding site. Topology properties (properties 9–14 in the Binding site column of Supplementary Table 1c) describe the local structure of the binding site.

The change in solvent accessible surface area (defined as $\Delta\mathrm{SASA}_{AB} = \mathrm{SASA}_A + \mathrm{SASA}_B - \mathrm{SASA}_{AB}$), number of residues and number of secondary structure elements describe the extent of the interface. The residue types present, secondary structure types present and change in polar solvent accessible surface area are more fine-grained properties describing the actual physical structures present and their chemical composition. Two measures of binding site continuity were computed. The number of structural patches was determined by counting the number of connected components in a graph representation of binding site residues where an edge was placed between residues within 6 Å of each other. The number of sequence segments was counted to describe the sequence continuity of each binding site. The continuity properties of the interfaces were calculated by summing the properties from their corresponding binding sites.

The number of residue pairs, number of secondary structure element contacts and number of interatomic contacts describe a combination of the extent and complexity of the interface. The residue contact types secondary structure contact types and secondary structure topology capture the nature and complexity of the interface. The interatomic contacts are further categorized into van der Waals contacts, hydrogen bonds salt bridges and disulfide bridges based on distance criteria [H-bond criteria as defined by JOY (Mizuguchi *et al.*, 1998); disulfide bridge defined when two Cys S atoms are closer than 3.0 Å].

As for the domain connectivity graphs, a crude linear representation of the secondary structure topology graph was generated for use as a topological fingerprint to group the structures into topological classes (Supplementary information: Topological Fingerprints, Supplementary Figure 2c and d).

Redundancy removal and clustering

Two types of clustering were performed on the domain–domain interfaces. The first procedure, redundancy removal, aims to provide a non-redundant set of interfaces for analysis by addressing the issue of duplicate interface structures. The second procedure, clustering, aims to group together similar interfaces to aide in the understanding of interface diversity. Although both procedures involve clustering, they serve different purposes.

Removal of redundancy of domain–domain interfaces

Redundancy, in the form of duplicate interface structures, exists for several reasons: redundancy within PDB entries, interfaces duplicated in derived PQS structures and redundancy across different PDB entries. The first two types of redundancy are explicitly addressed by hierarchically clustering interfaces associated with each PDB code using a distance function that combines the following properties: types of residue–residue contacts present (represented as a 210-bit vector, *aa*), buried solvent accessible surface area (Δ SASA) and the number of residues in the interface (numres) [Equation (1)]. The bit vectors were compared using the Hamann distance measure, $\mathrm{dist}_{\mathrm{hamann}}$, a rescaled and reversed version of the traditional Hamann similarity coefficient, $\mathrm{sim}_{\mathrm{hamann}}$, developed for use in plant systematics [Supplementary information: Equations (1) and (2)] (Hamann, 1961). The resulting dendrogram was cut into clusters using a strict threshold of 0.1. This cutoff corresponds to maximum differences of 10% in the buried surface areas, 10% in the numbers of residues or 0.1 Hamann distance in the residue–residue type contact vectors. The cluster membership of each interacting domain pair is stored in PIBASE. The clustering was performed on the interacting domain pairs list, prior to the buried surface area filter. This procedure identified ∼75% of the domain pairs as redundant (Table 1).

$$\mathrm{d}_{A,B} = \frac{1}{3}\left(\mathrm{dist}_{\mathrm{hamann}}(aa_A, aa_B) + \left(1 - \frac{\min(\Delta\mathrm{SASA}_A, \Delta\mathrm{SASA}_B)}{\max(\Delta\mathrm{SASA}_A, \Delta\mathrm{SASA}_B)}\right) + \left(1 - \frac{\min(\mathrm{numres}_A, \mathrm{numres}_B)}{\max(\mathrm{numres}_A, \mathrm{numres}_B)}\right)\right) \quad (1)$$

Table 1. PIBASE content

	SCOP	CATH
Structures		
Structures (PDB and PQS)		38 940
Associated PDB codes		20 740
Domains	120 110	103 246
Interfaces		
Interacting domain pairs	158 915	138 286
Interfaces (ΔSASA $\geq$ 300 Å^2)	86 127	76 746
Redundancy		
Interacting domain pairs unique within structure file	77 105	
Interacting domain pairs unique within PDB code	41 493	
Unique and ΔSASA $\geq$ 300 Å^2	20 912	

Unless otherwise noted, all the numbers shown represent data obtained from both PDB and PQS structures. The number of interacting domain pairs are shown using both SCOP and CATH definitions. The interface clustering was performed only on the SCOP pairs. The ΔSASA filter of 300 Å^2 removes ~45% of the interacting domain pairs. The redundancy removal procedure flags ~75% of the interfaces as redundant.

The third type of redundancy, duplicate interfaces across PDB entries, can also be addressed in a similar fashion, but is not yet implemented. While the minimal three-property set was found to be effective at recognizing interface similarity within a PDB file, a different and likely larger set of interface properties are required for a more general interface similarity measure. However, the choice of specific properties to use depends heavily on the definition of redundancy, and the intended application. As such, we leave this clustering up to the user, while providing the appropriate tools (e.g. properties, clustering algorithms).

Clustering of complexes, interfaces and binding sites

The topological fingerprints were used to group the non-redundant complexes, interfaces and binding sites into discrete topological classes. The complexes were grouped according to their domain connectivity (Supplementary Figure 2b), while the interfaces and binding sites were grouped according to the topology of their secondary structure elements (Supplementary Figure 2c). The clustering reveals 989, 18 755 and 30 975 topological groups of complexes, interfaces and binding sites, respectively.

In the current implementation, groups are formed by members with identical topological fingerprints. A more refined distance metric for topology fingerprints would be useful in describing a continuous gradation of topology similarity. However, such a clustering will depend on a specific application, and is therefore beyond the scope of the current paper.

Implementation

PIBASE was implemented using the MySQL relational database system (http://www.mysql.com). It was built by a set of Perl programs using the DBI interface to communicate with the MySQL system. Most properties were computed with MODELLER (Sali and Blundell, 1993). Secondary structure assignments were made by DSSP (Kabsch and Sander, 1983). Interatomic distances were computed using an inhouse ANSI C implementation, called kd-contacts, of the median kd-trees algorithm (Friedman *et al.*, 1977; Berg *et al.*, 1998). The kd-trees algorithm, a commonly used computational geometry algorithm, performs nearest neighbor queries by first building a tree in $O(n \log n)$ time, and then querying it in $O(n^{1-1/d} + k)$ time, where n is the number of data points in the d-dimensional space, and k is the number of reported points (Supplementary information: kd-trees algorithm). This approach is much faster than the naive approach of all versus all distance calculation [$O(n^2)$]. The logarithmic scaling allows rapid calculation of contact maps even for large structures with tens of thousands of atoms, such as PQS entries of virus capsids.

The clustering of the distance matrix for redundancy removal was performed using an inhouse Perl library. The calculations were done in a parallel fashion on 50 2.6 GHz Pentium IV processors in ~15 h. The database is updated automatically with every SCOP and CATH release.

Accessibility

PIBASE is accessible via the world wide web at http://salilab.org/pibase. The interface allows the user to query the database by PDB codes, complex topology fingerprints and domain classification codes. The range of possible queries will expand as users request additional functionality.

While a web interface is well suited for standard queries with relatively simple conditions, a programming interface can be more useful for complex queries. A Perl library, used in the construction of the database, will be released shortly, allowing complex queries to be performed without the complexity of directly accessing the underlying MySQL structures. In addition, the contents of the database tables, as well as a schema describing the logical relationships between the tables, are available for download.

Composition of PIBASE

Briefly, PIBASE currently contains 158 915 interacting domain pairs between 105 061 domains from 2125 SCOP families. More interface structures are available between domains from the same SCOP family (1405 homo-family pairs) than different SCOP families (982 hetero-family pairs) (Fig. 2a). Of a total of 2567 families in the SCOP classification, interface structures are available for 1946 of them.

Visually, it is obvious that the distribution of partner structural similarities is non-uniform (Fig. 2b). To investigate this distribution further, we compared the observed distribution of partner structural similarities to a random model in which all SCOP families interact with all other SCOP families. We found that interactions between domains with similarities only at the superfamily level are overrepresented (~5-fold). Interactions between domains from the same fold are almost twice as abundant as expected from the random model. Interactions between domains from the same SCOP class are approximately the same as expected. The interfaces between structurally dissimilar domains are underrepresented (~2-fold). In summary, the structures of interacting domain pairs currently available are weighted towards partners with the most structural similarity along the SCOP hierarchy. However, it is difficult to conclude from this observation that actual protein interaction networks behave in this manner, as the observed preferences likely reflect both an actual non-uniform distribution of structural similarity between interacting partners and sampling bias in the PDB.

Interfaces are observed to be mostly continuous in structure, but very segmented in sequence. On average, ~78 residues are presented to the interface on ~34 secondary structure elements, or ~23 continuous sequence segments (Fig. 3a, Supplementary

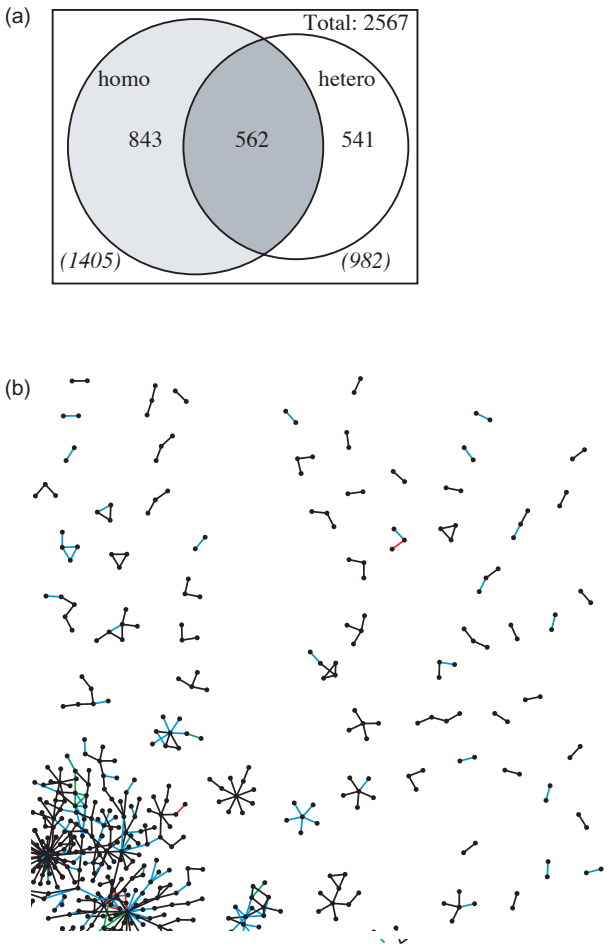

Fig. 2. Interactions in SCOP space. **(a)** Venn diagram of interaction coverage in SCOP space. Regular font represents the numbers of SCOP families. Italics represents the numbers of interfaces (e.g. 1405 homo-family SCOP family pairs). **(b)** Partial SCOP domain family interactome. Nodes represent SCOP families. Edges represent structurally observed interfaces. Edge color represents the SCOP similarity of the interacting nodes (Red = same superfamily, Green = same fold, Blue = same class, Black = no similarity). Only interfaces between different SCOP families are shown. Only a quarter of the full graph is shown. Graph layout by LGL (Adai *et al.*, 2004).

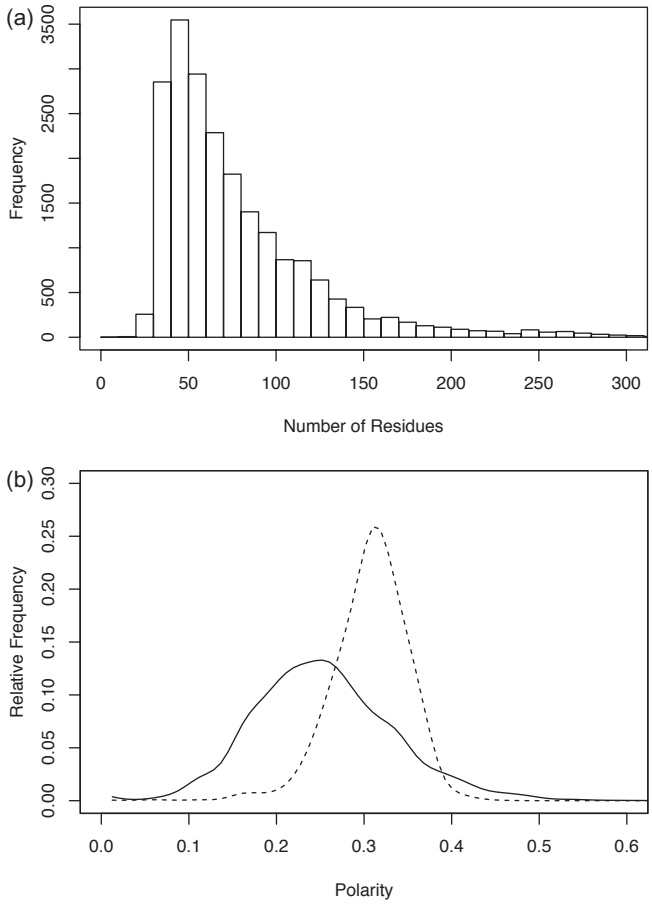

Fig. 3. PIBASE interface property distributions. All the plots are calculated from the non-redundant set of interfaces. **(a)** Number of residues at the interface. Maximum 862 residues (data not shown). **(b)** Interface polarity (solid) compared to whole domain surface polarity (dashed). The polarity of an interface is defined by the fraction of buried surface area contributed by polar atoms (N, O). Similarly, the polarity of the whole domain surface is defined by the fraction of solvent accessible surface area contributed by polar atoms. The whole domain surface polarities were calculated from all SCOP domains.

Figure 3a and b). However, each interface usually involves only two structurally continuous patches, one contributed by each binding site (Supplementary Figure 3c). As expected, the sequence discontinuity is directly proportional to the buried surface area of the interface ($r^2 = 0.71$, Supplementary Figure 3d).

An example of the physicochemical properties that can be analyzed using PIBASE is the polarity of interfaces (Fig. 3b). The polarity of each interface was defined by the fraction of the buried solvent accessible surface area that was contributed by polar atoms (N, O). The interface polarity exhibits a broad distribution with a mean of ~25%. This distribution is slightly more hydrophobic than the whole domain surface, which exhibits a narrower distribution with a mean of ~30%. The polarity of the whole domain surface was similarly defined, as the fraction of the solvent accessible surface area contributed by polar atoms.

DISCUSSION

We presented a comprehensive database of structurally characterized protein complexes. We first described its construction (Fig. 1), followed by its content (Table 1, Fig. 2). From domain topology to secondary structure topologies at individual interfaces, we presented groupings at size scales from the entire complex to individual interfaces. We also described the distributions of a subset of the interface properties stored in PIBASE (Fig. 3).

Several collections of protein chain and domain interfaces have been recently reported (Tsai *et al.*, 1996; Jones and Thornton, 1996; Hitz and Honig, 1999; Park *et al.*, 2001; Keskin *et al.*, 2004). A SCOP domain family interactome was published that supplemented SCOP interfaces extracted from the PDB with those observed in yeast protein interaction data (Park *et al.*, 2001). This resource allowed the proposal of possible evolutionary reasons for the observed repertoire of family–family intermolecular and intramolecular interactions. More recently, a collection has been created of non-redundant high-resolution structures of protein chain pairs extracted from the

PDB (Keskin *et al*., 2004). The interfaces were clustered using geometric hashing (Nussinov and Wolfson, 1991), a sequence order-independent structural superposition algorithm, which allowed the detection of conserved interface architectures across different fold types. The datasets reported vary widely in their size and breadth of descriptors, as expected from the different types of analysis they were designed for.

The main goals in developing PIBASE have been completeness of its domain interface coverage as well as diversity of the descriptors calculated at various scales. Though it contains a comprehensive set of interfaces, filters can easily be applied to focus on a specific type of interface or on those with a given minimum experimental resolution. The explicit topological clustering, previously developed for fold classification (West head *et al*., 1999), is unique in its application to protein complexes.

The completeness of PIBASE makes it suited for investigations into the structure of protein interactions, as well as for benchmarking methods such as protein-protein docking. To illustrate its utility as a general purpose bioinformatics resource, we list here several current applications in our group.

The interfaces stored in PIBASE have been used as templates for the prediction of protein interaction partners (Pieper *et al*., 2004). Candidate interaction partners are generated by detecting pairs of proteins from the same genome that contain domains for which an interface has been observed. These candidate interactions are then assessed by building comparative models of the individual proteins and scoring their putative interface using a statistical potential that captures residue type contact preferences at interfaces. This method predicts not only interaction partners, but also binding modes. Similar schemes have been previously reported (Aloy and Russell, 2002; Lu *et al*., 2003). The interaction predictions have been deposited in MODBASE (Pieper *et al*., 2004).

The spatial localization of protein binding sites in PIBASE has been analyzed (D.Korkin, F.P.Davis, A.Sali, manuscript in preparation). Localization is a measure that describes the degree of overlap of the binding sites observed for a given protein domain family. The lower the localization, the more scattered the distribution of binding sites. A range of localization values are observed for domain families. Many families exhibit a higher localization than expected by random (e.g. obligate homo-dimeric enzymes such as alkaline phosphatase), while others exhibit a lower localization than expected by random (e.g. highly divergent families such as C-type lectins).

The binding sites stored in PIBASE are also used by LS-SNP, a large-scale structural annotation of human single nucleotide polymorphisms (SNPs) (Karchin *et al*., 2005). This analysis combines multiple types of sequence and structure information, including protein binding sites, to predict whether an observed SNP is functionally deleterious.

Lastly, PIBASE has been integrated into an automated structure annotation system in the DBAli (Marti-Renom *et al*., 2001) and MODBASE (Pieper *et al*., 2004) databases. As structural genomics efforts are rapidly determining protein structures, it becomes important to annotate them using automated methods which leverage existing knowledge. For a given input protein structure, a structural alignment program MAMMOTH (Ortiz *et al*., 2002) is used to find similarities to known protein structures, and the SALIGN module of MODELLER (M.S.Madhusudhan, M.A.Marti-Renom, N.Eswar, A.Sali, manuscript in preparation) is applied to prepare multiple alignments of similar protein structures. The query protein structure can then inherit numerous properties from similar characterized structures, including ligand binding sites from LIGBASE (Stuart *et al*., 2002), and binding partners from PIBASE.

The modular and relational design of PIBASE allows easy cross-referencing to other databases of protein structure, sequence and function. Work is currently underway to cross-reference binary protein interaction databases such as BIND (Bader *et al*., 2003), using MODBASE structural annotation of the interacting proteins (Pieper *et al*., 2004). Through further integration with the plethora of high quality databases, PIBASE will become a valuable resource for the structural biology community.

ACKNOWLEDGEMENTS

We would like to thank the members of the Sali laboratory for valuable comments and suggestions, in particular Frank Alber, Maya Topf, Damien Devos, MS Madhusudhan, Mike F. Kim, Dmitri Korkin, Eswar Narayanan and Ursula Pieper. We acknowledge funding by NSF (EIA-0325004), as well as computer hardware gifts from Sun, Intel and IBM. F.P.D. acknowledges support from a Howard Hughes Medical Institute predoctoral fellowship.

REFERENCES

Adai,A.T. *et al*. (2004) LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *J. Mol. Biol*., **340**, 179–190.

Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.

Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci., USA*, **99**, 5896–5901.

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

Argos,P. (1988) An investigation of protein subunit and domain interfaces. *Protein Eng*., **2**, 101–113.

Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res*., **31**, 248–250.

Berg,M.D., Kreveld,M.V., Overmars,M. and Schwarzkopf,O. (1998) *Computational Geometry: Algorithms and Applications*, 2nd ed. Springer Verlag, Berlin.

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

Caffrey,D.R. *et al*. (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*., **13**, 190–202.

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

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

Editorial (2004) Psi-phase 1 and beyond. *Nat. Struct. Mol. Biol*., **11**, 201.

Elcock,A.H. and McCammon,J.A. (2001) Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl Acad. Sci., USA*, **98**, 2990–2994.

Fields,S. and Song,O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

Friedman,J.H. *et al*. (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, **3**, 209–226.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Hamann,U. (1961) Merkmalsbestand und verwandtschaftsbeziehungen der farinosae. Ein beitrag zum system der monokotyledonen. *Willdenowia*, **2**, 639–768.

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

Hitz,B. and Honig,B. (1999) Spin-PP: Surface properties of interfaces—protein–protein interfaces.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

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

Janin,J. *et al*. (1988) Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol*., **204**, 155–164.

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

Jones,S. *et al.* (2000) Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, **13**, 77–82.

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

Karchin,R., Kelly,L. and Sali,A. (2005) Improving functional annotation of nonsynonomous snps with information theory. *Pac. Symp. Biocomput*, 397–408.

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

Lu,L. *et al.* (2003) Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.*, **13**, 1146–1154.

Marti-Renom,M.A. *et al.* (2001) Dbali: a database of protein structure alignments. *Bioinformatics*, **17**, 746–747.

Mizuguchi,K. *et al.* (1998) Joy: protein sequence–structure representation and analysis. *Bioinformatics*, **14**, 617–623.

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.

Nussinov,R. and Wolfson,H.J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.

Ofran,Y. and Rost,B. (2003) Analysing six types of protein–protein interfaces. *J. Mol. Biol.*, **325**, 377–387.

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Ortiz,A.R. *et al.* (2002) Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Park,J. *et al.* (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, **307**, 929–938.

Pieper,U. *et al.* (2004) Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, 32 (Database issue), D217–D222.

Ponting,C.P. and Russell,R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 45–71.

Richmond,T.J. and Richards,F.M. (1978) Packing of alpha-helices: geometrical constraints and contact areas. *J. Mol. Biol.*, **119**, 537–555.

Robert,C.H. and Ho,P.S. (1995) Significance of bound water to local chain conformations in protein crystals. *Proc. Natl Acad. Sci. USA*, **92**, 7600–7604.

Robert,C.H. and Janin,J. (1998) A soft, mean-field potential derived from crystal contacts for predicting protein–protein interactions. *J. Mol. Biol.*, **283**, 1037–1047.

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

Sali,A. (2003) NIH workshop on structural proteomics of biological complexes. *Structure (Camb)*, **11**, 1043–1047.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sali,A. and Overington,J.P. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.*, **3**, 1582–1596.

Sali,A. *et al.* (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.

Stuart,A.C. *et al.* (2002) Ligbase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, **18**, 200–201.

Tsai,C.J. *et al.* (1996) A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Valdar,W.S. and Thornton,J.M. (2001) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.

Veretnik,S. *et al.* (2004) Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.*, **339**, 647–678.

Westhead,D.R. *et al.* (1999) Protein structural topology: automated analysis and diagrammatic representation. *Protein Sci.*, **8**, 897–904.