Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D501–D506***
doi:10.1093/nar/gkr1128

DOMMINO: a database of macromolecular interactions

Xingyan Kuang[1], Jing Ginger Han[1], Nan Zhao[1], Bin Pang[1], Chi-Ren Shyu[1] and Dmitry Korkin[1,2,*]

[1]Informatics Institute and Department of Computer Science and [2]Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA

Received August 15, 2011; Revised October 10, 2011; Accepted November 8, 2011

ABSTRACT

With the growing number of experimentally resolved structures of macromolecular complexes, it becomes clear that the interactions that involve protein structures are mediated not only by the protein domains, but also by various non-structured regions, such as interdomain linkers, or terminal sequences. Here, we present DOMMINO (http://dommino.org), a comprehensive database of macromolecular interactions that includes the interactions between protein domains, interdomain linkers, N- and C-terminal regions and protein peptides. The database complements SCOP domain annotations with domain predictions by SUPERFAMILY and is automatically updated every week. The database interface is designed to provide the user with a three-stage pipeline to study macromolecular interactions: (i) a flexible search that can include a PDB ID, type of interaction, SCOP family of interacting proteins, organism name, interaction keyword and a minimal threshold on the number of contact pairs; (ii) visualization of subunit interaction network, where the user can investigate the types of interactions within a macromolecular assembly; and (iii) visualization of an interface structure between any pair of the interacting subunits, where the user can highlight several different types of residues within the interfaces as well as study the structure of the corresponding binary complex of subunits.

INTRODUCTION

Interactions mediated by proteins play a crucial role in many cellular processes (1). An important step toward a mechanistic description of these processes is the structural analysis and modeling of these interactions and their interaction interfaces. Today, experimental and computational methods can routinely provide an increasing number of high-resolution structures of the binary and higher order complexes (2). These complexes are collected into databases, the interactions that formed the complexes are grouped based on their structures and functions and their important features are analyzed.

Currently, there are a number of open access structural databases containing structure of the protein–protein interactions and/or interaction interfaces extracted from RCSB Protein Data Bank [PDB, (3)] and Protein Quaternary Server [PQS, (4)]. Among the most popular and recent databases are 3DComplex (5), 3DID (6), DOMINE (7), iPfam (8), PIBASE (9), PSIBASE (10), SCOPPI (11), SNAPPI-DB (12). Most of the databases focus on the interactions between the protein domains. To define domains they employ either sequence-based or structure-based domain classification definitions, such as SCOP (13), CATH (14) and PFAM (15). Often, the interactions are further classified based on structural or physico-chemical features of their interfaces or based on the function of the corresponding interaction partners. For instance, SCOPPI classifies the interactions within each SCOP family according to the geometric features of the interface structures (11). Another database, 3DID uses gene ontology functional annotation to characterize the extracted domain–domain (6). Finally, some databases, such as PIBASE and 3DID, have the ability to visualize the network of interacting proteins or their domains. The reliance on the structural classification of proteins when characterizing an interaction is important, however, the protein interaction databases that employ structural classifications are often dependent on how frequent a classification definition is updated. As a result, many interaction databases are updated on the annual basis or less frequently, while the current rate of protein-mediated

*To whom correspondence should be addressed. Tel: +1 573 882 4762; Fax: +1 573 882 8318; Email: korkin@korkinlab.org

complexes growth in PDB is hundreds of structures each month.

Another common feature of many structural databases on protein interactions is that they focus on the interactions between either protein domains or single-domain proteins. While this covers a significant part of protein-mediated macromolecular interactions, other types of interactions exist. There are a few databases that focus on the interactions that occur between a protein and a peptide (16,17). For instance, a database with a name similar to ours, DOMINO, curates >3900 such interactions from literatures and is able to display the interaction networks (16). Another database, PepX, stores 1431 representations of PDB entries containing protein–peptide interactions and clusters them into 505 clusters (17).

In addition to the interactions mediated by protein domains and peptides, the interactions mediated by the unstructured protein regions such as C- and N-termini, as well as interdomain linkers are often considered as independent functional (and structural) regions of a protein, since they have been known to carry out important protein functions (18–20). For instance, the functions of N-termini include ubiquitination-induced degradation and protein sorting, while C-termini are associated with signaling and translation termination (19). Most importantly, the protein termini have been often found to interact with other proteins, with the examples ranging from the interactions between C-termini and multidomain scaffold proteins (21) to the interactions between N-termini and integral membrane proteins (22). The other functionally important regions of the multidomain proteins, interdomain linkers, are often less conserved than the neighboring domains and have been found to underlie a wide range of functions (20,23). The less-studied linker–protein interactions have been associated with the regulation of kinase activity (24) and allosteric communication (25). The datasets of individual unstructured regions have been collected and analyzed (26–29), however, a systematic collection and analysis of the structures of their interactions is yet to be done.

Here, we introduce DOMMINO, a Database Of Macro-Molecular INteractiOns that currently houses >500 000 binary interactions mediated by proteins. By analysis of the existing databases, several important features have been integrated into DOMMINO. First, it is fully automated is designed to update itself on a weekly basis, followed by the weekly PDB update. Second, it has an expanded coverage of types of interactions: in addition to storing domain–domain and domain–peptide interactions traditionally considered by the macromolecular interaction databases, DOMMINO includes the interactions mediated by the unstructured regions, such as interdomain linkers, N-terminal and C-terminal regions. Third, the database has an expanded coverage of structural domains by integrating the manual annotation of protein domains participating in the interactions using the latest version of SCOP (13) with the automated annotation using SUPERFAMILY (30). Finally, the web-interface is designed to provide the researcher with a pipeline to study the interactions: from a flexible search to the subunit interaction networks to the individual interaction interfaces.

METHODS

There are six types of subunits that form interactions in DOMMINO: a protein domain, interdomain linker (or simply linker), C-terminal region, N-terminal region, peptide and an undefined chain. The subunit types are defined, and the binary interactions are annotated via the following three stages of data processing (Figure 1). First, we determine all protein domains and extract their coordinates by combining domain definitions manually curated by SCOP and domain predictions by a Hidden Markov Model based approach. Second, using the above domain annotations we determine all interdomain linkers and terminal regions of the proteins and extract their coordinates. Finally, we extract the coordinates of protein peptides and undefined protein chains.

Data sources and preprocessing

DOMMINO integrates the information from multiple sources. The structural data on macromolecular interactions are extracted from PDB (3); the atomic coordinates are processed (ATOM and HETATM records) for each PDB structure with at least one protein chain. If a PDB entry has more than one structure model, the first model is used in the database's current implementation. For domain assignment, the most recent release (June 2009) of manually curated SCOP database is used that includes the manual annotation for 38 221 PDB entries. The SCOP domain definitions are extracted from file dir.cla.scop.txt. During the domain assignment, each PDB structure is assigned to one of two groups as follows. If a PDB structure has at least one assigned SCOP protein domain, according to the SCOP definition file, the structure is assigned to the first group, otherwise it is assigned to the second group of macromolecular complexes for which the constituting domains are later predicted using SUPERFAMILY software (30) ('Domain annotation' section).

Domain annotation

Out of 110 800 SCOP protein domain definitions extracted from file dir.cla.scop.txt, a slightly reduced set of 109 942 definitions is employed for the first group of PDB structures, because some SCOP domains cannot be located in the coordinate records of the PDB files from the current PDB release. The definitions are stored as a new parsable file filtered.cla.scop.txt, in which the same data format as in dir.cla.scop.txt is used.

To assign protein domains for the PDB structures from the second group, for which no SCOP annotation is available, SUPERFAMILY software is used (30). The software is designed to accurately predict the SCOP domains based on a sequence of each protein chain. It employs a collection of hidden Markov models (HMMs) each corresponding to a structural protein domain at the SCOP family level. The prediction is done by scanning a protein sequence against the HMMs and is accepted when

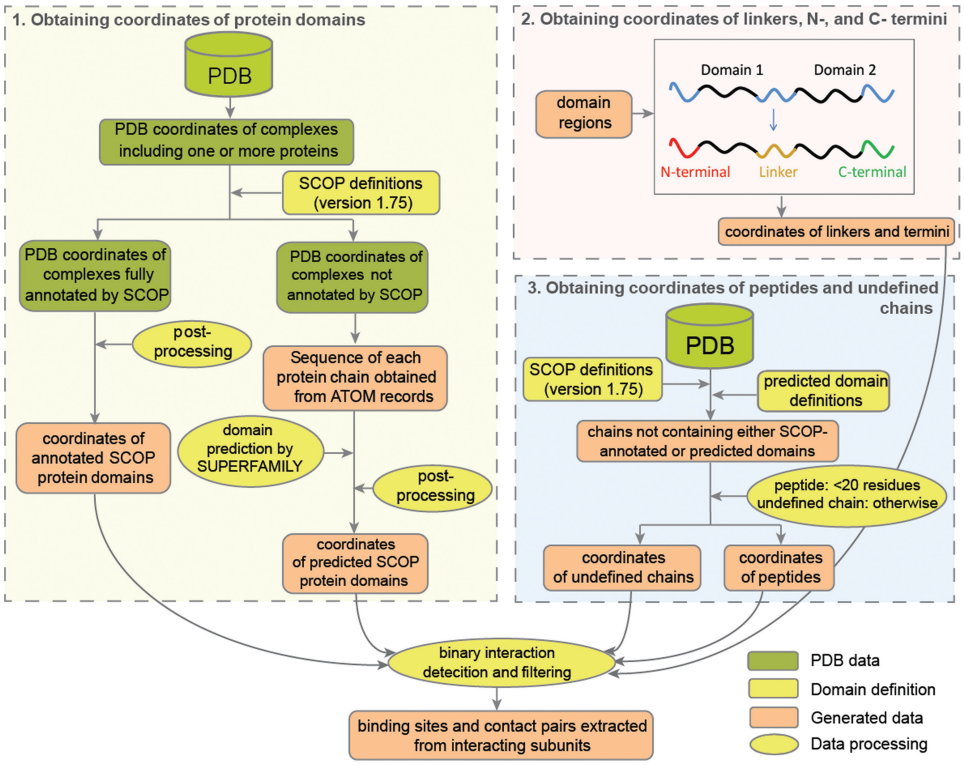

Figure 1. Data processing in DOMMINO. Data processing consists of three stages: (**1**) determining protein domains, (**2**) determining inter-domain linkers and terminal protein regions and (**3**) determining protein peptides and unlabeled protein chains. During each stage, the coordinates of subunits of each type are collected and the corresponding interactions are annotated.

E-value $\leq$0.01. In the current version of DOMMINO, we do not predict domains spanning multiple chains in a PDB structure. When two predicted domains overlap, we evenly distribute the overlapping region between the two predicted domains. As a result, 101 389 domains were predicted for the PDB files from the second group, and their coordinates were extracted. The information on predicted domains, including the corresponding regions and the SCOP classification, is summarized in a parsable file pre.cla.scop.txt.

Locating linkers, C-terminal and N-terminal regions

A predicted SCOP domain can be either a whole protein chain or its fragment. A domain annotated by SCOP can also consist of more than one fragment of a protein chain. Each protein fragment that does not belong to any protein domain is then annotated as a linker, C- or N-terminus, depending on whether it is surrounded by two domain fragments or there is just one domain fragment located to the right or to the left of it, correspondingly. For each defined region, its coordinates are also extracted.

Determining peptides and undefined chains

There are chains in PDB files that cannot be annotated by assigning domain using either SCOP definitions or prediction by SUPERFAMILY. Based on how long such protein chains, they are classified as either peptides or undefined chains. Specifically, we use a 20-residue threshold to determine a peptide (if the number of residues is $<$20) and undefined chain (otherwise). The same threshold has been used before in ASTRAL, a similar domain definition protocol (31).

Determining protein interfaces

For each PDB structure, each pair of determined subunits of the above six types is analyzed to determine whether they interact with each other using the following definition. If any atom of a residue in one protein subunit is within 6 Å of any atom of a residue in another protein subunit, the two residues are determined as the contact pair residues. Each of the residues belongs to the corresponding protein binding site for the respective subunit. When determining an intrachain interaction between two protein subunits that are sequentially adjacent to each other, the distance of the C-terminal residue of the first subunit can be potentially closer than 6 Å to the N-terminal residue of the second subunit, while these residues do not contribute to the interaction. To exclude such cases, a pair of residues from the same chain but belonging to different adjacent subunits is not recorded as a contact pair if these residues are $<$10 residues apart. Finally, DOMMINO allows the user to apply different threshold levels on the number of contact pairs to exclude potential artifactual interactions (the default threshold is 10 contact pairs).

DATABASE CONTENT AND MAINTENANCE

The content of DOMMINO includes both intrachain and interchain interactions between a pair of subunits of any of the six types. After the current update (August, 2011) the database has ~514 000 entries (that is, all binary interactions sharing at least one contact pair), ~146 000 of which are determined using the SCOP domain definitions and ~368 000 using domain predictions by SUPERFAMILY. The database has larger number of interchain (~57.3%) than intrachain interactions. Approximately 47.7% of all interactions are domain–domain interactions, with the interactions between a domain and a C-terminus being the second most populated type of interaction (~18.5%) and the interactions between a domain and a peptide being the third most populated type (~17.2%). Interestingly, all 21 possible types of interactions in DOMMINO are populated with the interaction structures. Combined, the interactions mediated by the unstructured protein regions (i.e. linkers and termini) comprise ~47.6% of all interactions. There are 8751 (~1.7%) interactions between two different species. The interaction data in DOMMINO is organized as a relational database that supports the web-based search functions.

In order to be fully synchronized with the current release of PDB, DOMMINO and the corresponding file system are automatically updated every week following the PDB's weekly updating schedule. During the weekly update of PDB, some older entries could be deleted from the current release as erroneous or outdated while, new entries could be added. The PDB log file that records the deleted entries and new added entries is used to first flag the deleted entries in DOMMINO. Then, the new added PDB entries are processed. If a PDB ID of new added PDB entry does not exist in DOMMINO, the entire three-stage processing ('Methods' section) is applied for this PDB entry to obtain the interaction data. When the PDB entry is present in the database, we remove the old PDB entry first and all related data from DOMMINO, before adding the entry with the same PDB ID.

USER INTERFACE

A web-based interface that relies on the DOMMINO is developed for investigating protein-mediated interactions (Figure 2A). The web-based interface includes four basic features: (i) search query, (ii) retrieval of results, (iii) visualization of the subunit interaction network for a PDB structure and (iv) visualization of the structure either of the interaction complex or of the interaction interface for a selected binary interaction. All these stages are integrated together into a single workflow to facilitate the structure-based analysis of protein-mediated interactions.

Search query and PDB list of result

The user can retrieve data in three different ways: (i) basic search, (ii) advanced search and (iii) data browsing and downloading. When the user aims to investigate a protein interaction for a specific PDB entry, they can choose the simple search option and enter entering a PDB ID. In simple search, the threshold on the minimal number of contact pairs is set to 10. The advanced search option makes the search more flexible using four criteria options including interaction type (e.g. domain to domain, domain to linker, undefined chain to C-terminal), SCOP family, organism name and keyword. The keyword includes a description of a target protein molecule, e.g. 'PEPTIDE ARYLATION ENZYME'. The user can combine these criteria creating a more complex and therefore more focused search query. In advanced search, the user can setup a different contact pair threshold. For data browsing, we list all possible interaction types. If the user selects a type of interaction mediated by at least one domain (e.g. D–D or D–C), they will be prompted to choose a SCOP family for the interacting domain(s). Otherwise, the web server retrieves the entire PDB ID list for the selected type of interactions. When the user submits a search query, the web server will retrieve a list of PDB entries containing the queried protein interactions (see Figure 2B, for an example).

Interaction network

The web-based search tool of DOMMINO provides the means to visualize the interaction network between the subunits of a PDB entry from the retrieved results (see Figure 2C, for an example of the interaction network for PDB ID 2HTK). In the network includes all subunits from a PDB entry, irrespective whether or not they participate in the interactions. Edges are used to connect two interacting subunits. Six geometric shapes are implemented to represent the six types of protein subunits. Subunit nodes from the same chain are represented with the same color. The label in a subunit node indicates the chain this subunit belongs to, and the sequential order of the subunit in the chain. The complete description of the shapes and colors used in the network can be retrieved by clicking the help button just below the network image.

Structure visualization

By selecting two interacting subunits from the subunit interaction network, the user can visualize the structure of the complex and its binding sites or select to show the interaction interface only, using a Jmol-based tool (32). The Jmol-based web tool provides four molecular representation options including default (ball-and-stick), ribbon, backbone and surface representations to view the structure of the interaction complex, its interfaces, or its binding sites (Figure 2D). In addition, the user can highlight by different colors the important physicochemical properties of the residues, such as hydrophobicity, charge, etc.

DISCUSSION

In this work, we presented a new database that combines the interactions between the structural subunits of six different types. The database interface provides the user with a three-stage pipeline to study macromolecular interactions including a flexible search, visualization of subunit interaction network, as well as visualization of an interface structure between any pair of the interacting subunits. In

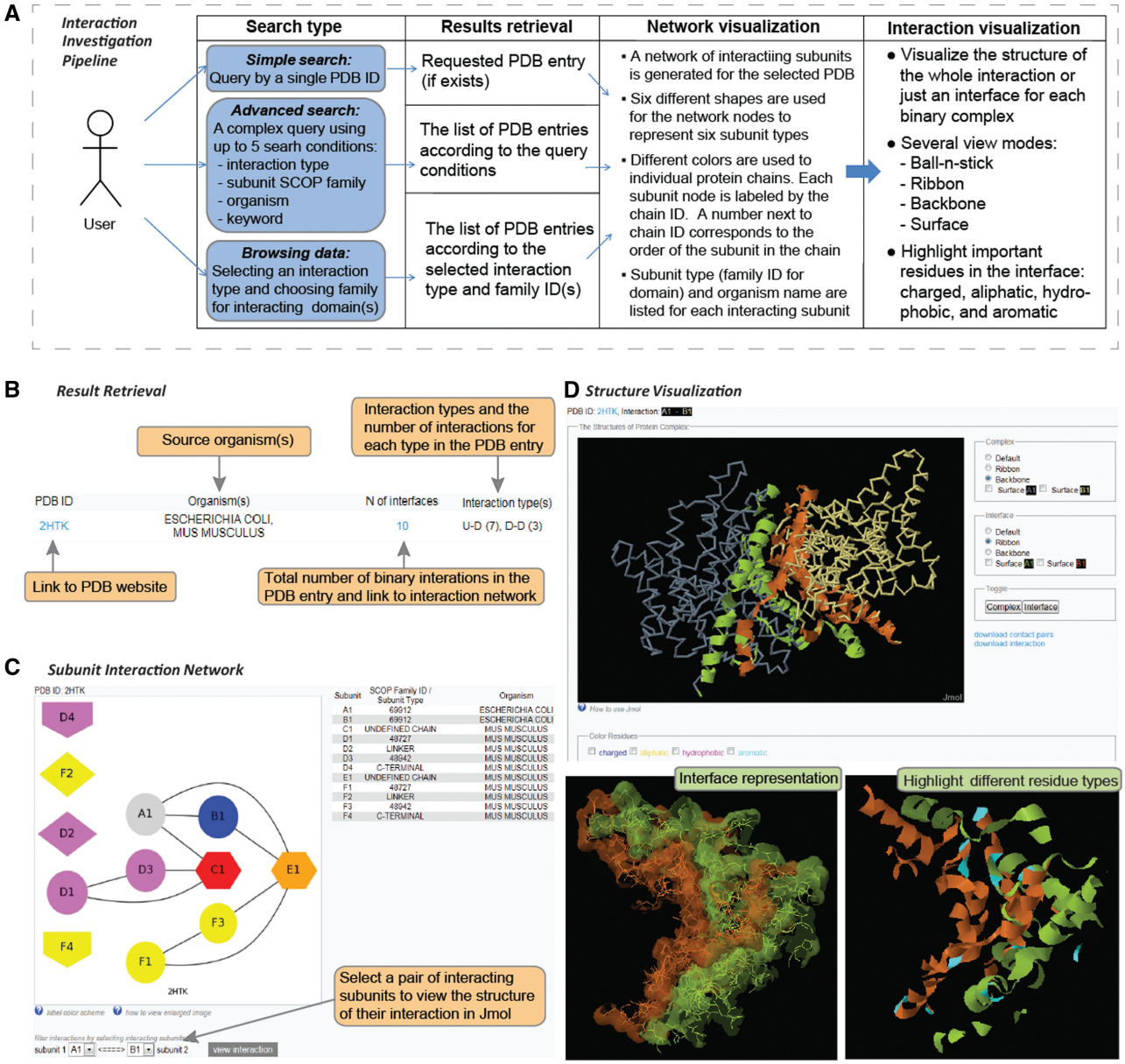

Figure 2. User interface of DOMMINO. (**A**) A typical workflow to study macromolecular interactions in DOMMINO through its web-interface. (**B**) An example of retrieved information for a simple query using PDB ID 2HTK. (**C**) A subunit interaction network visualized for the above PDB structure. (**D**) An interaction between subunits A1 and B1 from the same PDB structure. Shown are three basic ways to visualize the interaction: (i) as a whole complex, (ii) as an interface with shown accessible surface or (iii) as an interface with highlighted interface residues of different types.

addition to its weekly updates, DOMMINO will automatically update the SCOP domain definitions when the new definitions become available. Our next future step is to include to DOMMINO the interactions mediated by other macromolecules, such as DNAs and RNAs.

ACKNOWLEDGEMENTS

The authors are grateful to the developers and research groups who assisted with their software packages and databases including Jmol package, the SUPERFAMILY package, SCOP database and the Protein Data Bank.

FUNDING

National Science Foundation (DBI-0845196 to D.K.); Paul K. and Diane Shumaker Endowment (fellowships to X.K. and N.Z.). Funding for open access charge: National Science Foundation (DBI-0845196).

Conflict of interest statement. None declared.

REFERENCES

1. Alberts,B. (1998) *Essential Cell Biology : An Introduction to the Molecular Biology of the Cell*. Garland Pub., New York.

2. Alber,F., Forster,F., Korkin,D., Topf,M. and Sali,A. (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.*, **77**, 443–477.
3. Berman,H.M. (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr. A*, **64**, 88–95.
4. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
5. Levy,E.D., Pereira-Leal,J.B., Chothia,C. and Teichmann,S.A. (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, e155.
6. Stein,A., Ceol,A. and Aloy,P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
7. Yellaboina,S., Tasneem,A., Zaykin,D.V., Raghavachari,B. and Jothi,R. (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.*, **39**, D730–D735.
8. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
9. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
10. Gong,S., Yoon,G., Jang,I., Bolser,D., Dafas,P., Schroeder,M., Choi,H., Cho,Y., Han,K., Lee,S. *et al.* (2005) PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*, **21**, 2541–2543.
11. Winter,C., Henschel,A., Kim,W.K. and Schroeder,M. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.
12. Jefferson,E.R., Walsh,T.P., Roberts,T.J. and Barton,G.J. (2007) SNAPPI-DB: a database and API of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Res.*, **35**, D580–D589.
13. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
14. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. and Orengo,C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
15. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
16. Ceol,A., Chatr-aryamontri,A., Santonico,E., Sacco,R., Castagnoli,L. and Cesareni,G. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res.*, **35**, D557–D560.
17. Vanhee,P., Reumers,J., Stricher,F., Baeten,L., Serrano,L., Schymkowitz,J. and Rousseau,F. (2010) PepX: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Res.*, **38**, D545–D551.
18. Hayes,C.S., Alarcon-Hernandez,E. and Setlow,P. (2001) N-terminal amino acid residues mediate protein-protein interactions between DNA-bound alpha /beta -type small, acid-soluble spore proteins from Bacillus species. *J. Biol. Chem.*, **276**, 2267–2275.
19. Chung,J.J., Shikano,S., Hanyu,Y. and Li,M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends Cell. Biol.*, **12**, 146–150.
20. Gokhale,R.S. and Khosla,C. (2000) Role of linkers in communication between protein modules. *Curr. Opin. Chem. Biol.*, **4**, 22–27.
21. Stricker,N.L., Christopherson,K.S., Yi,B.A., Schatz,P.J., Raab,R.W., Dawes,G., Bassett,D.E. Jr, Bredt,D.S. and Li,M. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. *Nat. Biotechnol.*, **15**, 336–342.
22. Siniossoglou,S. and Pelham,H.R. (2002) Vps51p links the VFT complex to the SNARE Tlg1p. *J. Biol. Chem.*, **277**, 48318–48324.
23. Gokhale,R.S., Tsuji,S.Y., Cane,D.E. and Khosla,C. (1999) Dissecting and exploiting intermodular communication in polyketide synthases. *Science*, **284**, 482–485.
24. Briggs,S.D. and Smithgall,T.E. (1999) SH2-kinase linker mutations release Hck tyrosine kinase and transforming activities in Rat-2 fibroblasts. *J. Biol. Chem.*, **274**, 26579–26583.
25. Zhuravleva,A. and Gierasch,L.M. (2011) Allosteric signal transmission in the nucleotide-binding domain of 70-kDa heat shock protein (Hsp70) molecular chaperones. *Proc. Natl Acad. Sci. USA*, **108**, 6987–6992.
26. George,R.A. and Heringa,J. (2002) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.*, **15**, 871–879.
27. Bahir,I. and Linial,M. (2005) ProTeus: identifying signatures in protein termini. *Nucleic Acids Res.*, **33**, W277–W280.
28. Carugo,O. (2011) Participation of protein sequence termini in crystal contacts. *Protein Sci.*, **20**, 2121–2124.
29. Lange,P.F. and Overall,C.M. (2011) TopFIND, a knowledgebase linking protein termini with function. *Nat. Methods*, **8**, 703–704.
30. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
31. Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.*, **30**, 260–263.
32. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.