JAIL: a structure-based interface library for macromolecules

Stefan Günther[1], Joachim von Eichborn[1], Patrick May[2] and Robert Preissner[1,*]

[1]Institute of Molecular Biology and Bioinformatics, Charité-University Medicine Berlin, Arnimallee 22, 14195 Berlin and [2]Max-Planck-Institute for Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

ABSTRACT

The increasing number of solved macromolecules provides a solid number of 3D interfaces, if all types of molecular contacts are being considered. JAIL annotates three different kinds of macromolecular interfaces, those between interacting protein domains, interfaces of different protein chains and interfaces between proteins and nucleic acids. This results in a total number of about 184 000 database entries. All the interfaces can easily be identified by a detailed search form or by a hierarchical tree that describes the protein domain architectures classified by the SCOP database. Visual inspection of the interfaces is possible via an interactive protein viewer. Furthermore, large scale analyses are supported by an implemented sequential and by a structural clustering. Similar interfaces as well as non-redundant interfaces can be easily picked out. Additionally, the sequential conservation of binding sites was also included in the database and is retrievable via Jmol. A comprehensive download section allows the composition of representative data sets with user defined parameters. The huge data set in combination with various search options allow a comprehensive view on all interfaces between macromolecules included in the Protein Data Bank (PDB). The download of the data sets supports numerous further investigations in macromolecular recognition. JAIL is publicly available at http://bioinformatics.charite.de/jail.

INTRODUCTION

Proteins interact quickly and specifically with each other or with nucleic acids. All interactions form a biochemical network that reflects the high complexity of cellular metabolism. Nevertheless, the vast majority of all interactions are not yet identified and are subject to current research (1). An important step towards the mechanistic descriptions of such interactions is the 3D structural information of macromolecules. However, complexed proteins are difficult to co-crystallize and the number of publicly available X-ray structures in the Protein Data Bank (PDB) (2) is very limited (3). Thus, a structure-based analysis of particular interacting macromolecules is often only possible by using docking models. For systematic analyses the problem of the low number of protein–protein complexes might be avoided by taking into consideration interacting domains or chains. Such types of interfaces often exhibit a similar behaviour like those of interacting proteins (4). For instance, knowledge-based potential functions that represent the co-occurrence of certain residues might be similar for contacts between domains of a single chain as well as contacts between proteins. Another approach is the utilization of the interfaces between domains or chains to detect structural similarities to binding sites of interacting proteins. First applications apply this method for protein–protein complex modelling (5). Consequentially, some structure-based databases exist that focus on the interacting parts of proteins. SCOPPI (6), SNAPPI (7) and PIBASE (8) classify interfaces between domains, the domain information was retrieved from SCOP (9), CATH (10) or Pfam (11). Since they depend on domain definitions extracted from secondary databases especially structures solved during the last few years are normally not yet classified (12). HotSprint (13) focuses on conserved residues in chain contact sites and is regularly updated but domain information is ignored. Dockground (3) comprises the so far most comprehensive data set of interacting proteins and chains respectively. The excellent database also provides user defined data sets of the associated unbound protein-binding sites. Nevertheless, it focuses on protein–protein interaction, so information about intra-chain contacts is not retrievable. None of the mentioned databases contains interfaces between proteins and nucleic acids. Although each application is useful to enlighten questions the database is specialized for, a comprehensive web resource that combines all the different kinds of interfaces between macromolecules is not available. Furthermore, the database should be characterized by regular updates and the

*To whom correspondence should be addressed. Tel: +49-30-8445-1649; Fax: +49-30-8445-1551; Email: robert.preissner@charite.de

Figure 1. Example for a search by homology. (**a**) Entry of PDB-ID 2FR4. The highlighted link (Show related proteins) yields a list of homologous proteins. (**b**) List of homologs of 2FR4 and the associated interfaces. The last entry is PDB-ID 2J88 and is shown in Figure 2b.

opportunity to download appropriate data sets. To overcome this lack we developed JAIL, a structure-based interface library for macromolecules.

DATABASE

Currently, the database contains more than 184 000 interfaces that are composed of four different fractions: 81 000 interfaces between domains classified by SCOP, 76 000 interfaces between different protein chains, 8000 interfaces between proteins and nucleic acids and 19 000 interfaces which were calculated based on the assumed biological units. Since they were not directly solved cristallographically, they are annotated separately. The interfaces result from the evaluation of 52 000 different asymmetric unit files as well as the associated biological unit files provided by the PDB. An interface is defined as those atoms of a chain or domain that are located within a range of 10 Å around the Cα-atoms of the interacting counterpart. In the case of nucleic acids the backbone (P/C4′)-atoms were considered. Each binding site has to consist of at least five Cα-atoms as the case may be

backbone atoms of the nucleic acids. Assumed biological units were calculated based on the first two models build up by reflection of the unit cells. Information on the evolutionary conservation was extracted from the HSSP database (14). The PDB-IDs of nucleic acids containing structures were retrieved from the Nucleic Acid Database (15). All chains of the database were sequentially clustered using the regular updated lists provided by the PDB calculated by the Cd-hit program (16). Thus, it is possible, to select interfaces of proteins, which are similar in sequence to each other as well as to download non-redundant data sets based on protein sequences. Structural clustering was implemented by the selection of representative interfaces of each family–family or superfamily–superfamily contact between domains classified in SCOP. A protein can be identified by a detailed search form that allows searches by PDB-ID, protein name, EC-number, UniProt accession number or SCOP-ID. An implemented full text search allows the screening to the full header information of the structure-file as well as the SCOP domain descriptions. Visualization of the interfaces was implemented by pre-generated thumbnails of each interface and on the other hand by the interactive protein visualizer Jmol
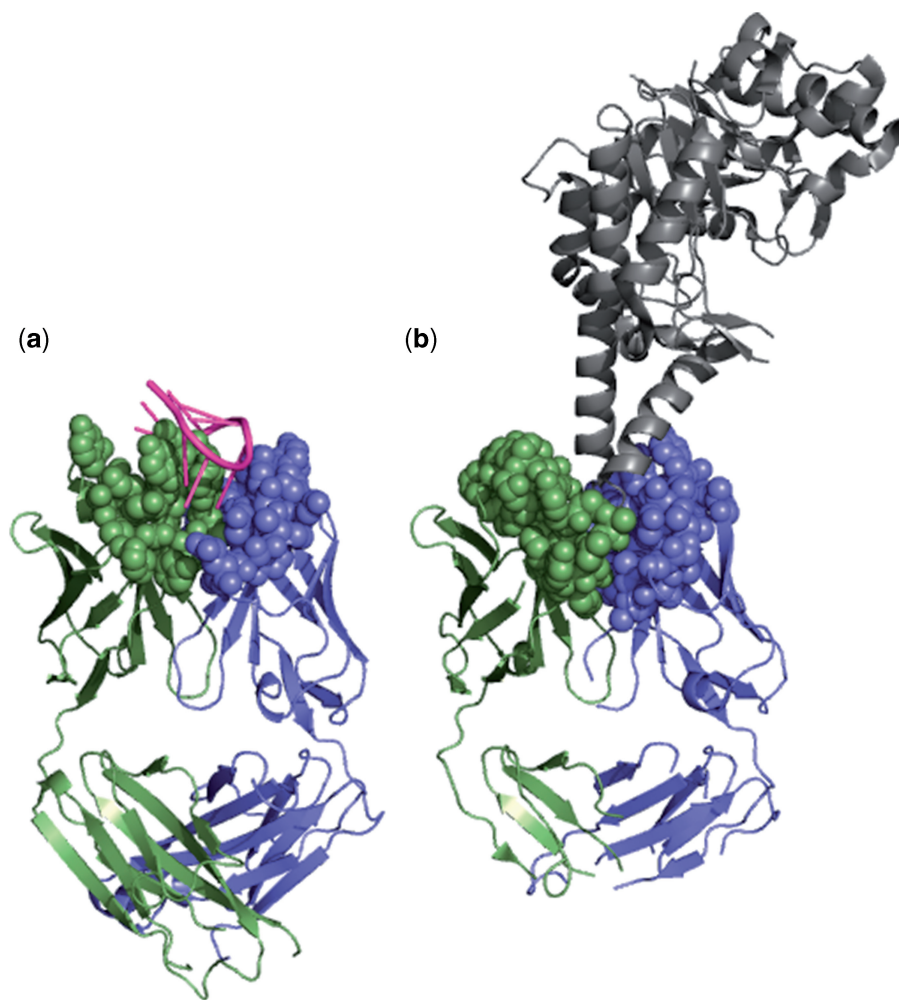
Figure 2. Two similar Fab-fragments (sequence identity >95%) which interact with different kinds of macromolecules. The complexes were identified with the homology search option (see Figure 1) of JAIL. (**a**) Fab-fragment in complex with a stem-loop DNA (PDB-ID: 2FR4). (**b**) A monoclonal IgG Fab-fragment in complex with hyaloronidase (PDB-ID: 2J88).

(http://www.jmol.org). The download section provides various possibilities to build up a user defined data set. Structurally, clustered interfaces based on the SCOP domain definitions as well as sequentially clustered interfaces based on protein chain clustering are separately retrievable but can also be combined. Parameters like the SCOP-hierarchy (family/superfamily) or the sequence identity level (50%, 70%, 90% and 95%) are selectable. The database is automatically updated six times a year.

EXAMPLE OF USE

The provided browsable interfaces support the answering of various further investigations. One of them is the comparative study of molecular recognition of nucleic acids and proteins. For instance, experimental evidence exists, that proteins mimic nucleic acids to usurp the role of interacting macromolecules (17,18). Shape comparisons of different kinds of interfaces (protein–protein/protein–nucleic acids) may help to identify cases of molecular mimicry.

A matter of particular interest in this context is the identification of protein domains that interact with other proteins as well as with nucleic acids. Figure 1 shows an example of a search for such a case by using the implemented 'Show related proteins'-option. Figure 2 shows the resulting interfaces of a fab-fragment in complex with an enzyme. The same domain architecture is also capable to bind single stranded hairpin DNA.

REFERENCES

1. Amaral,L.A. (2008) A truer measure of our ignorance. *Proc. Natl Acad. Sci. USA*, **105**, 6795–6796.
2. Berman,H.M., Bhat,T.N., Bourne,P.E., Feng,Z., Gilliland,G., Weissig,H. and Westbrook,J. (2000) The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol.*, **7 (Suppl)**, 957–959.
3. Gao,Y., Douguet,D., Tovchigrechko,A. and Vakser,I.A. (2007) DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. *Proteins*, **69**, 845–851.
4. Tuncbag,N., Gursoy,A., Guney,E., Nussinov,R. and Keskin,O. (2008) Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.*, **381**, 785–802.
5. Gunther,S., May,P., Hoppe,A., Frommel,C. and Preissner,R. (2007) Docking without docking: ISEARCH–prediction of interactions using known interfaces. *Proteins*, **69**, 839–844.
6. 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.
7. 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.
8. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
9. 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.
10. Greene,L.H., Lewis,T.E., Addou,S., Cuff,A., Dallman,T., Dibley,M., Redfern,O., Pearl,F., Nambudiry,R., Reid,A. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.
11. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
12. Rother,K., Michalsky,E. and Leser,U. (2005) How well are protein structures annotated in secondary databases? *Proteins*, **60**, 571–576.
13. Guney,E., Tuncbag,N., Keskin,O. and Gursoy,A. (2008) HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **36**, D662–D666.
14. Dodge,C., Schneider,R. and Sander,C. (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.*, **26**, 313–315.
15. Berman,H.M., Westbrook,J., Feng,Z., Iype,L., Schneider,B. and Zardecki,C. (2002) The Nucleic Acid Database. *Acta. Crystallogr. D Biol. Crystallogr.*, **58**, 889–898.
16. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
17. Liang,H. and Landweber,L.F. (2005) Molecular mimicry: quantitative methods to study structural similarity between protein and RNA. *Rna*, **11**, 1167–1172.
18. Putnam,C.D. and Tainer,J.A. (2005) Protein mimicry of DNA and pathway regulation. *DNA Repair (Amst)*, **4**, 1410–1420.