BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 6 2005, pages 827–828
doi:10.1093/bioinformatics/bti098

Databases and ontologies

PDZBase: a protein–protein interaction database for PDZ-domains

Thijs Beuming[1], Lucy Skrabanek[1,2], Masha Y. Niv[1], Piali Mukherjee[2] and Harel Weinstein[1,2,*]

[1]Department of Physiology and Biophysics, and [2]Institute for Computational Biomedicine, Weill Medical College of Cornell University, 1300 York Ave., New York, NY 10021, USA

Received on August 14, 2004; revised on October 5, 2004; accepted on October 8, 2004
Advance Access publication October 28, 2004

ABSTRACT

Summary: PDZBase is a database that aims to contain all known PDZ-domain-mediated protein–protein interactions. Currently, PDZBase contains approximately 300 such interactions, which have been manually extracted from >200 articles. The database can be queried through both sequence motif and keyword-based searches, and the sequences of interacting proteins can be visually inspected through alignments (for the comparison of several interactions), or as residue-based diagrams including schematic secondary structure information (for individual complexes).

Availability: http://icb.med.cornell.edu/services/pdz/start.

Contact: pdzbase@med.cornell.edu.

INTRODUCTION

PDZ (PSD-95, Discs-large, ZO-1) domains are ubiquitous protein–protein interaction domains comprising about 70–90 residues (Nourry *et al*., 2003; Hung and Sheng, 2002; Fan and Zhang, 2002; Sheng and Sala, 2001; van Ham and Hendriks, 2003). They are involved in numerous interactions with various proteins, in a variety of biological processes. Some proteins contain multiple copies of PDZ-domains, often in combination with other protein–protein interaction domains. This architecture enables simultaneous interactions among several proteins, thus turning the PDZ-domain-containing proteins into putative 'molecular switchboards' (Dueber *et al*., 2003). The most prominent role of PDZ-containing proteins appears to be the assembly of protein complexes at the plasma membrane, where they bind to the C-termini of membrane proteins.

The specificity of PDZ-domain-based interactions is determined primarily by the sequence of the C-terminus of the proteins they bind. Thus, the specificity of PDZ-interactions has been traditionally attributed to the last three residues of the ligand (i.e. positions P-0, P-1, P-2, counting backwards from the terminal residue in the ligand). A classification of PDZ-binding motifs in the C-termini has been proposed, in which the consensus sequence for class I is S/T–X–Φ, and for class II is Φ–X–Φ (where Φ is any hydrophobic residue), with the corresponding PDZ-domain classified into class I or class II binding (Songyang *et al*., 1997). More recently, this classification system has been challenged based on the discovery that some PDZ-binding sequences do not belong to either of the two classes and the observation that certain PDZ-domains promiscuously bind to both class I and class II ligands. Moreover, it has become apparent that residues further N-terminal are important for specificity as well. Indeed, the human genome alone contains hundreds of PDZ-domains that are able to bind specific targets, a feat which would seem difficult to achieve with as few as three or four (relatively similar) recognition sites. A comprehensive comparative analysis of a large number of PDZ-peptide complexes is expected to yield further insights into determinants of specificity. Such a task, and other types of comparative analyses of the important PDZ-interactions will require an easily accessible source of specialized data, given the very large number of complexes between PDZ-domains and ligands that have been reported. To construct such a source of data, we have systematically analyzed all articles in the PubMed database containing the keyword PDZ (>1000 articles) and have constructed a comprehensive database to store the known interactions. A web-based server, named PDZBase, that allows for querying and analyzing of this database of PDZ-domain-mediated protein–protein interactions, is presented below.

DATABASE CONTENT

PDZBase currently contains ∼300 interactions, all of which have been manually extracted from the literature, and have been independently verified by two curators. The extracted information comes from *in vivo* (co-immunoprecipitation) or *in vitro* experiments (GST-fusion or related pull-down experiments). Interactions identified solely from high throughput methods (e.g. yeast two-hybrid or mass spectrometry) were not included in PDZBase. Other prerequisites for inclusion in the database are: (1) that knowledge of the binding sites on both interacting proteins must be available (for instance through a truncation or mutagenesis experiment); (2) that interactions must be mediated directly by the PDZ-domain, and not by any other possible domain within the protein. The database is continuously maintained and will be updated regularly with new interactions reported in the literature.

IMPLEMENTATION AND DATABASE SCHEMA

For PDZBase, we have used the Java Data Object API (JDO) to connect between the web application logic layer and the database backend. This technology presents the strong advantage that persistent objects are modeled as object-oriented classes (in the Java

*To whom correspondence should be addressed.

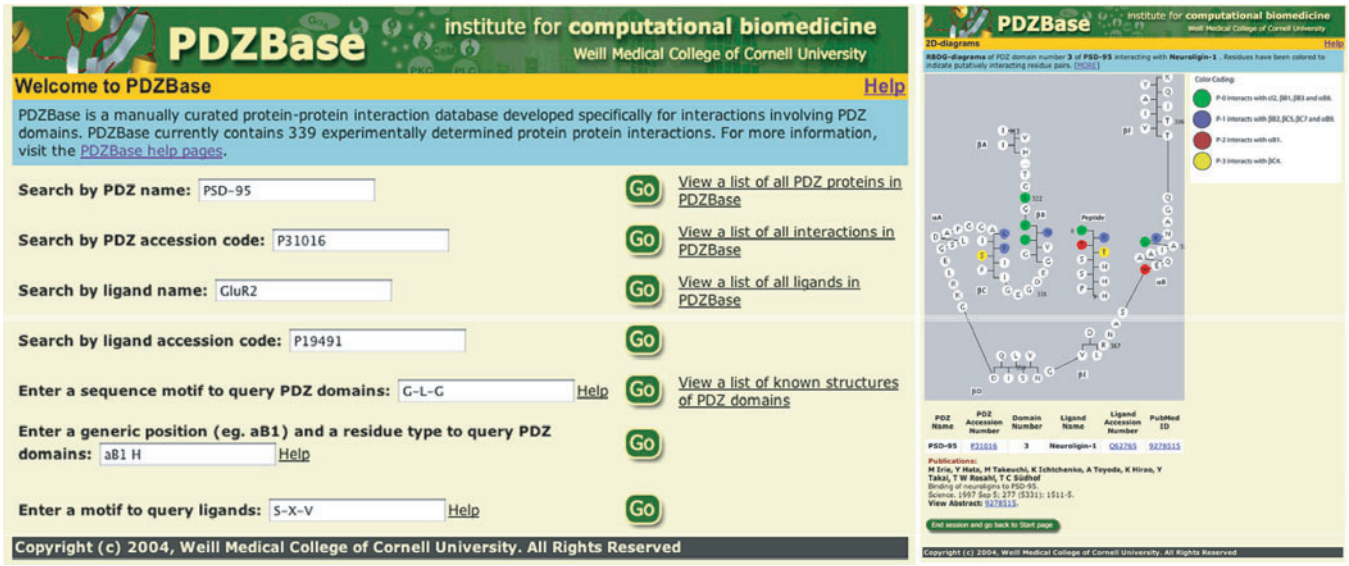

Fig. 1. PDZBase interface. Proteins or interactions can be queried using names, identifiers, or motifs (left). Details of interactions are represented as 2D-diagrams (right).

language), but can be stored either in a relational DBMS or an object-oriented DBMS. The Kodo implementation of JDO has been used to connect to an Oracle 8.1.7 backend. The JDOQL is used to query the database.

The database structure schema consists of the classes PossibleInteraction, PDZProtein, PDZDomain and Ligand. The PossibleInteraction class contains the PDZProtein, the interacting PDZDomain, the Ligand, the literature reference and information about the location of the interface. To permit storage of non-existing interactions (negative controls) as well, a Phenotype field in PossibleInteraction can indicate whether the interaction exists or not. The PDZProtein and Ligand classes contain the Swiss-Prot accession code (Boeckmann *et al*., 2003), the amino acid sequence of the protein and the organism. The PDZDomain class contains the start and end points of the domain, and its location within the protein (i.e. which domain number).

DATABASE ACCESS

PDZBase currently provides a simple search interface (Fig. 1, left) that enables the database to be queried for interactions using the names and external identifiers (e.g. Swiss-Prot; Boeckmann *et al*., 2003) of the interacting proteins. Additionally, sets of proteins in the database can be retrieved if they have a specific sequence motif in common, or share a specific residue type at a certain position. For instance, querying with 'S-X-V' in the 'Enter a motif to query ligands' search field (Fig. 1) returns all ligands in PDZBase with a Ser at P-2 and a Val at P-0. A query with 'aB1 H' in the 'Enter a generic position and a residue type to query PDZ domains' search field returns all PDZ-domains with a His at position αB1 (the first residue of the second α-helix). All interactions involving the chosen subset of proteins can then be retrieved. The sequences of interacting proteins can be visualized as an HTML-formatted alignment, and exported as a FASTA format text-file. Finally, each interaction is linked to a details-page, which shows the residues of the interacting proteins on a 2D-diagram generated by the residue-based-diagram-generator (RbDg; Campagne *et al*., 2003) (Fig. 1, right), and provides interaction-specific links to other databases (Swiss-Prot; Boeckmann *et al*., 2003, PubMed).

ACKNOWLEDGEMENTS

We thank Dr Nathalie Basdevant for curating part of the interactions. The work is supported by NIH grants K05 DA00060, P01 DA124080 and P01 DA12923.

REFERENCES

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Campagne,F., Bettler,E., Vriend,G. and Weinstein,H. (2003) Batch mode generation of residue-based diagrams of proteins. *Bioinformatics*, **19**, 1854-1855.

Dueber,J.E., Yeh,B.J., Chak,K. and Lim,W.A. (2003) Reprogramming control of an allosteric signaling switch through modular recombination. *Science*. **301**, 1904–1908.

Fan,J.S. and Zhang,M. (2002) Signaling complex organization by PDZ domain proteins. *Neurosignals*, **11**, 315–321.

Hung,A.Y. and Sheng,M. (2002) PDZ domains: structural modules for protein complex assembly. *J. Biol. Chem.*, **277**, 5699–5702.

Nourry,C., Grant,S.G. and Borg,J.P. (2003) PDZ domain proteins: plug and play! *Sci. STKE*, **2003**, RE7.

Sheng,M. and Sala,C. (2001) PDZ domains and the organization of supramolecular complexes. *Annu. Rev. Neurosci.*, **24**, 1–29.

Songyang,Z., Fanning,A.S., Fu,C., Xu,J., Marfatia,S.M., Chishti,A.H., Crompton,A., Chan,A.C., Anderson,J.M. and Cantley,L.C. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science*, **275**, 73–77.

van Ham,M. and Hendriks,W. (2003) PDZ domains–glue and guide. *Mol. Biol. Rep.*, **30**, 69–82.