BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 1 2002
Pages 200–201

LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures

*Ashley C. Stuart, Valentin A. Ilyin and Andrej Sali**

Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA

Received on June 8, 2001; revised on August 3, 2001; accepted on August 3, 2001

ABSTRACT

Summary: A database comprising all ligand-binding sites of known structure aligned with all related protein sequences and structures is described. Currently, the database contains approximately 50 000 ligand-binding sites for small molecules found in the Protein Data Bank (PDB). The structure–structure alignments are obtained by the Combinatorial Extension (CE) program (Shindyalov and Bourne, *Protein Eng.*, **11**, 739–747, 1998) and sequence–structure alignments are extracted from the ModBase database of comparative protein structure models for all known protein sequences (Sanchez *et al.*, *Nucleic Acids Res.*, **28**, 250–253, 2000). It is possible to search for binding sites in LigBase by a variety of criteria. LigBase reports summarize ligand data including relevant structural information from the PDB file, such as ligand type and size, and contain links to all related protein sequences in the TrEMBL database. Residues in the binding sites are graphically depicted for comparison with other structurally defined family members. LigBase provides a resource for the analysis of families of related binding sites.

Availability: LigBase is accessible on the web at http://guitar.rockefeller.edu/ligbase.

Contact: ash@guitar.rockefeller.edu; sali@rockefeller.edu

The genome projects are generating sequence data at an astounding pace. However, the true prize of genomics will come from understanding the functions of these new proteins. Although protein function is best determined through extensive characterization in the laboratory, many potential functions may be gleaned from a computational comparison of an uncharacterized protein with previously characterized proteins (Koonin *et al.*, 1998). Minimally, such comparisons may suggest experiments required to define the function of a protein with less effort. While experimental analyses are always preferable to predictions, the scale of the genomes currently overwhelms the available resources for characterization by experiment. Thus, comparative analyses *in silico* provide an efficient method of discovery for uncharacterized proteins.

In addition to detection of sequence similarity across whole protein domains, it is useful to search for short sequence patterns indicative of functional sites. Although sequence pattern searches, such as Prosite (Hofmann *et al.*, 1999), are helpful in finding common binding or active sites, understanding the importance of particular residues comes from knowledge of the residues' location in space with respect to the ligand. In addition, structural aspects of a binding site can be conserved even when the sequences diverge. Combining sequence and structural information in a database provides a better tool to study ligand-binding sites than inspection of sequence conservation alone. Indeed, binding sites with a partial sequence motif or no motif may be missed entirely with sequence pattern searches, but can be revealed in structural models of the proteins in question (Fetrow *et al.*, 1999; Wallace *et al.*, 1996). Several databases of 3D structures of ligand-protein complexes exist including RELIBASE (Hendlich, 1998) and Dictionary of Homologous Superfamilies (DHS) (Bray *et al.*, 2000); however, the LigBase database introduced in this communication provides a unique venue to compare known binding sites as well as potential binding sites in comparative models, especially when accessed through ModView. A comparison of known binding sites in experimentally determined protein structures and comparative models will make it easier to analyze the nuances of binding.

LigBase contains approximately 50 000 binding sites for the ligands found in approximately 14 000 experimentally determined protein structures in the Protein Data Bank (PDB). Amino acid residues with at least one atom within 5 Å of a ligand atom define the ligand binding site. Ligands that are closer than 1.8 Å to any protein atom are designated as covalently bound (cov), otherwise they are designated as noncovalently bound (nco). HETATM records in the PDB files are used to define the ligands

*To whom correspondence should be addressed.

themselves; however, water molecules are not retained in the database. Ligands consist of small molecules such as nucleotides, cofactors, drugs, metal ions, fatty acids, carbohydrates, etc.; nucleic acid and protein ligands are excluded. Schematic diagrams of the binding sites are generated with the Ligplot program (Wallace *et al.*, 1995). The pairwise structural alignments are obtained using the CE program (Shindyalov and Bourne, 1998). The relationships stored in LigBase are limited to CE alignments with a Z-score of 3.8 or higher, so that biologically relevant similarities are generally retained and random alignments are generally excluded. Multiple alignments are generated on the fly from the pairwise alignment series. Binding site residues for a given ligand are mapped onto the structural alignments and can be visualized using the ModView plugin (http://guitar.rockefeller.edu/modview), a sequence/structure 3D viewer (for Linux and IRIX platforms) (Ilyin *et al.*, submitted). Alignments may be selected for all chains that a ligand contacts so that complete information is retained for binding sites defined by multiple chains.

The database offers a web-based tool to analyze and/or reference ligands and ligand binding sites that are found in the PDB. The primary search page permits searches by PDB code, ligand code, protein or ligand description (keyword), experimental method, or the number of ligand atoms. A flag may also be set to select noncovalent ligands, covalent ligands (closer than 1.8 Å to a protein atom), or all ligands. The search results page provides information about the ligand in question as well as links to the PDB header file (Berman *et al.*, 2000), all known protein sequences from the TrEMBL database that were modeled on this PDB template in ModBase (Sanchez *et al.*, 2000), and to the CATH page (Orengo *et al.*, 1997) for the PDB structure. Links are also provided to a Ligplot diagram of the binding site and to a 'select alignments' page for comparison to related sequences based on structural alignments. The 'select alignments' page contains several filters for restricting the search results, similar to filters in CE (Shindyalov and Bourne, 2001). Additionally, chains with the same ligand bound may be selected.

The 'Get Alignments' button displays the active site residues mapped onto the structural alignments. Multiple alignments are shown for the ligand binding sequence(s) in question (in red and purple) as well as any additional sequences that were selected (in gold and blue). In all cases, the highlighted residues represent the position of binding residues for the selected ligand. The presence, absence or mutation of binding site residues can be seen in the alignment.

Future developments will involve a more complete integration of LigBase with ModBase, so that comparative models can be automatically flagged for potential binding sites and ligands. This integration will also dramatically increase the pool of available structures for the study of binding sites. LigBase is updated automatically to reflect the growth of the PDB and CE databases.

ACKNOWLEDGEMENTS

A.C.S. acknowledges the support of the Alfred P. Sloan foundation for a postdoctoral fellowship in Computational Biology. A.S. was an Alfred P. Sloan Research Fellow and is an Irma T. Hirschl Trust Career Scientist. The investigations have also been aided by grants from NIH (GM 54762), Mathers Foundation, and Merck Genome Research Institute.

REFERENCES

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bray,J.E., Todd,A.E., Pearl,F.M., Thornton,J.M. and Orengo,C.A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.*, **13**, 153–165.

Fetrow,J.S., Siew,N. and Skolnick,J. (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.*, **13**, 1866–1874.

Hendlich,M. (1998) Databases for protein–ligand complexes. *Acta Crystallogr. D. Biol. Crystallogr.*, **54**, 1178–1182.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Ilyin,V.A., Pieper,U., Stuart,A.C., Martí-Remon,M.A. and Sali,A. (2001) Visualization and analysis of multiple protein seqences and structures by ModView, submitted.

Koonin,E.V., Tatusov,R.L. and Galperin,M.Y. (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.*, **8**, 355–363.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Sanchez,R., Pieper,U., Mirkovic,N., de Bakker,P.I., Wittenstein,E. and Sali,A. (2000) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **28**, 250–253.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Shindyalov,I.N. and Bourne,P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the combinatorial extension (CE) algorithm. *Nucleic Acids Res.*, **29**, 228–229.

Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.

Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.*, **5**, 1001–1013.