BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 6 2004, pages 986–988
DOI: 10.1093/bioinformatics/bth048

PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt

Andrew C.R. Martin*

School of Animal and Microbial Science, The University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK

Received on July 29, 2003; revised on October 2, 2003; accepted on November 15, 2003
Advance Access publication February 5, 2004

ABSTRACT

Summary: A mapping between chains in the Protein Databank and Enzyme Classification numbers is invaluable for research into structure–function relationships. Mapping at the chain level is a non-trivial problem and we present an automatically updated Web-server, which provides this link in a queryable form and as a downloadable XML or flat file.

Availability: The query interface and downloadable files may be accessed at http://www.bioinf.org.uk/pdbsprotec/

Contact: andrew@bioinf.org.uk

Supplementary information: Further details of the methods used are available at http://www.bioinf.org.uk/pdbsprotec/methods/

1 INTRODUCTION

Assignment of Enzyme Classification (EC) numbers to Protein Databank (PDB; Berman *et al.*, 2000) chains allows functional assignments to be made to enzyme chains in the PDB. Such a mapping has been used in the analysis of protein fold distributions for different enzyme classes (Martin *et al.*, 1998) and has application in the automated predictive annotation of protein sequences assigned to a given fold by methods such as threading (Jones *et al.*, 1992). It can also be used as an aid to verifying protein structure classification schemes such as CATH (Orengo *et al.*, 1997) and SCOP (Murzin *et al.*, 1995), and for the general exploration of correlations between fold and function.

Assignment of EC numbers to PDB chains is a non-trivial task. Primarily, this is because PDB files generally only contain the EC number in the COMPND record which applies to the whole file. For example, in entry 3HFL (Sheriff *et al.*, 1987), the EC number 3.2.1.17 is found. Since this is the antibody HyHEL-5 (chains L and H) bound to its antigen, lysozyme (chain Y), the EC number should only be assigned to chain Y. The problem has been resolved in new PDB files, which have structured COMPND records that list MOL_ID and CHAINs and map the EC number(s) to each chain [e.g. entry 1KR1 (Bokma *et al.*, 2002)].

Previously, we mapped PDB chain to EC number to analyze the distribution of CATH folds among different enzyme classes (Martin *et al.*, 1998). Each chain was mapped to a SwissProt entry (Boeckmann *et al.*, 2003) and information from the Enzyme database (Bairoch, 2000) and SwissProt entries was used to obtain EC numbers. Mapping from PDB chain to SwissProt is non-trivial as cross-links from PDB to SwissProt, where present, can contain either the SwissProt identifier (ID) or the accession code (AC), but usefully are presented at the chain level, while links in the other direction are at the whole PDB file level, but are updated more frequently. FASTA (Pearson and Lipman, 1988) was used to resolve which chain was involved and a brute-force FASTA search against SwissProt was used for any unassigned chains. The whole process took some 3–4 days to run and unfortunately was not designed in a manner that could allow fast easy updates as new PDB or SwissProt entries become available.

The problem of PDB chain to SwissProt mapping has been partially addressed by the Research Collaboratory for Structural Bioinformatics (RCSB). Their beta site provides a mapping of PDB files to associated SwissProt entries (ftp://beta.rcsb.org/pub/pdb/uniformity/derived_data/pdb2sp.txt), but this simply lists all the entries associated with a file and gives no information about which chain is involved. The new XML format files containing PDB data (available at ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML/) now provide a mapping between 'entities' and SwissProt entries where an entity corresponds to a unique gene segment. This does appear to provide a reliable source of mappings, but extracting the EC number together with the associated chain and residue range on a regular automated basis remains a fairly complex task. In addition, the format of these files is still in a state of flux and the RCSB's alpha data have relevant differences. Like the original PDB files, both the flat file and the XML files still seem to suffer from inconsistencies in the use of SwissProt identifiers and accession codes.

The problem, however, has now been addressed by the Macromolecular Structure Database (MSD) group at the European Bioinformatics Institute (EBI). They now provide a

*Present address: Department of Biochemistry and Molecular Biology, Darwin Building, University College London, Gower Street, London WCIE 6BT.

mapping between PDB chain and SwissProt accession code at the individual residue level. This has the added benefit of correctly handling chimeric chains such as chain A of PDB entry 1GK5 (Chamberlin *et al.*, 2001). At the time of writing this mapping is updated on request, but should be available by FTP in an automatically updated form in the near future.

2 METHODS

From the EBI's PDB/SwissProt mapping we now obtain a simplified representation—a residue range within a PDB chain and the SwissProt accession code to which it maps. These data are imported into a table in a relational database implemented using PostgreSQL (http://www.postgresql.org/).

Second, we use the Enzyme database to obtain mappings between SwissProt codes and EC numbers. Additional mappings are then obtained from SwissProt, which is updated more regularly than the enzyme database and also contains partial assignments (e.g. EC number 1.1.1.-). These mappings are loaded into a second database table.

Queries of the resulting database allow PDB chains to be linked to SwissProt codes and thence to EC numbers. The complete mappings are exported from the database as a flat file and in XML. A Web query interface is provided, allowing queries on the basis of PDB code, SwissProt accession number or EC number. The results provide links to CATH (Orengo *et al.*, 1997), PDBSum (Laskowski, 2001) and the original SwissProt data via SRS (Etzold and Argos, 1993) at the EBI.

Central to the design of the system is automated updating. The Unix 'make' utility is used to update the database in a completely automated fashion. A similar approach was taken by us previously (Allcorn and Martin, 2002). The three data sources (the PDB/SwissProt mapping, the Enzyme database and SwissProt with all updates and modifications applied) are mirrored locally using the Perl 'mirror' script (http://sunsite.org.uk/packages/mirror/). 'Make' is then used to detect when these have been updated. If either the PDB/SwissProt mapping or the EC database is updated, the respective table is dropped and reloaded with the new data. If SwissProt is updated, then any additional information it supplies is added to the SwissProt/EC mapping table. The 'make' utility is run automatically on a nightly basis.

The Web site provides a search page allowing the database to be searched by PDB code (optionally with chain name), SwissProt accession code or EC number. Partial EC numbers may be entered (e.g. 1.1.1 will return all the 1.1.1.* entries). From the search page, the mapping between PDB residue range, SwissProt accession code and EC number may be downloaded as a flat file or an XML file.

The results page from a search displays the description of the relevant enzyme reaction followed by a table containing the results. This contains the PDB code, chain label (both of which are clickable to limit the search to this PDB code and chain respectively) and residue range, together with links to the CATH (Orengo *et al.*, 1997) and PDBSum (Laskowski, 2001) databases. The SwissProt accession code is provided together with a link to the SwissProt entry via SRS (Etzold and Argos, 1993) at the EBI. The accession code is clickable refining the search to show all entries for that SwissProt entry. Finally, the EC number (again clickable to refine the search) is displayed together with a link which provides a more detailed description of the enzyme reaction extracted from the Enzyme database.

Refinement of the search via the Web page is a useful facility for exploring relationships between proteins in the PDB. For example, suppose one searches initially for 3HFL (Sheriff *et al.*, 1987), an antibody bound to hen egg lysozyme. The results from this search include the SwissProt accession code 'P00698'. Clicking this provides a list of all the chains in the PDB which are structures for hen egg lysozyme. The search also provides the EC code (3.2.1.17) and clicking this provides a list of all lysozyme structures.

3 DISCUSSION

PDBSprotEC provides a reliable mapping between PDB chain and EC number in a totally automated fashion. As source data are updated, the database will be brought up to date with no user intervention.

The interface provides a searchable and browsable system. For example, one can enter a PDB code and identify its EC number. By clicking on that EC number, all PDB entries having the same EC number will be displayed with their SwissProt accession codes.

Other resources provide a partial mapping between PDB code and EC number. For example, PDBSum (Laskowski, 2001) extracts the EC number from the PDB file, but this is on a per-entry basis rather than per-chain. Similarly, the 'Enzyme Structures Database' (http://www.biochem.ucl.ac.uk/bsm/enzymes/), provides a browsable EC number-based index into the PDB, but suffers from the same per-entry problem and does not correctly handle chains with multiple EC numbers [e.g. entry 1KR1 (Bokma *et al.*, 2002)] as only the first EC number is stored (PDBSum lists them all). In neither case is the mapping downloadable for use in further analysis. In future, we hope to integrate this resource more closely with CATH and PDBSum.

ACKNOWLEDGEMENTS

The author thanks members of the MSD group at the EBI (Kim Henrick and Phil McNeil) for making the PDB/SwissProt mapping available.

REFERENCES

Allcorn,L.C. and Martin,A.C.R. (2002) SACS—a self-maintaining database of antibody crystal structures. *Bioinformatics*, **18**, 175–181.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

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.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., 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.

Bokma,E., Rozeboom,H.J., Sibbald,M., Dijkstra,B.W. and Beintema,J.J. (2002) Expression and characterization of active site mutants of hevamine, a chitinase from the rubber tree *Hevea brasiliensis*. *Eur. J. Biochem.*, **269**, 893–901.

Chamberlin,S.G., Brennan,L., Puddicombe,S.M., Davies,D.E. and Turner,D.L. (2001) Solution structure of the mEGF/TGFα44-50 chimeric growth factor. *Eur. J. Biochem.*, **268**, 6247–6255.

Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature (London)*, **358**, 86–89.

Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.

Martin,A.C.R., Orengo,C.A., Hutchinson,E.G., Jones,S., Karmirantzou,M., Laskowski,R.A., Mitchell,J.B.O., Taroni,C. and Thornton,J.M. (1998) Protein folds and functions. *Structure*, **6**, 875–884.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

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.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci., USA*, **85**, 2444–2448.

Sheriff,S., Silverton,E.W., Padlan,E.A., Cohen,G.H., Smith-Gill,S.J., Finzel,B.C. and Davies,D.R. (1987) Three-dimensional structure of an antibody–antigen complex. *Proc. Natl Acad. Sci., USA*, **84**, 8075–8079.