Application Note

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

*To whom correspondence should be addressed.

Running Head: Linking PDB chains to EC numbers

23rd July 2003

Abstract

Summary:
A mapping between chains in the Protein Databank (PDB) and Enzyme Classification (EC) 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 –or– a.c.r.martin@reading.ac.uk
Supplementary Information:
Further details of the methods used are available at `http://www.bioinf.org.uk/pdbsprotec/methods`

1 Introduction

The assignment of Enzyme Classification (EC) numbers to Protein Databank (Berman *et al.*, 2000, PDB) 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.

Unfortunately assignment of EC numbers to PDB chains is a non-trivial task. Primarilly this is because PDB files generally only contain the EC number in the COMPND record which applies to the whole file. Thus taking entry 3HFL (Sheriff *et al.*, 1987) as an example, the EC number 3.2.1.17 is found and appears to relate to the whole PDB entry. In reality this is the antibody HyHEL-5 (chains L and H) bound to its antigen, lysozyme (chain Y) and the EC number should only be assigned to chain Y.

Some time ago, we performed a mapping between PDB chain and EC number in order to analyze the distribution of CATH folds amongs different enzyme classes (Martin *et al.*, 1998). The method used was to map each chain to a SwissProt entry (Boeckmann *et al.*, 2003) and thence to use EC numbers appearing in the SwissProt entry and SwissProt codes appearing in the Enzyme database (Bairoch, 2000) to provide the mapping from PDB chain to EC number. Mapping from PDB entry to SwissProt code was, at the time, itself a non-trivial task. Cross-links from PDB to SwissProt, where present, could contain either the SwissProt identifier (ID) or the accession code (AC), but usefully are presented at the chain level. Links in the other direction are at the whole PDB file level, but are now updated more frequently than the PDB entries.

At that time we designed a protocol to perform the mapping which made use of the links from PDB to SwissProt and the reverse links from SwissProt to PDB then used sequence alignment with FASTA (Pearson and Lipman, 1988) to assign which specific chain was involved. Finally mappings between SwissProt codes and EC numbers were extracted first from the Enzyme database and then from SwissProt. The whole process takes 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 now been sidestepped by the Macromolecular Structure Database (MSD) group at the European Bioinformatics Institute (EBI) under the leadership of Kim Henrick, who now provide a mapping between PDB chain and SwissProt entry, not just at the whole-chain level but 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) which is a chimera of epidermal growth factor and human transforming growth factor alpha. At the time of writing this mapping is available on request, but should be available by FTP in an automatically updated form in the near future.

2 Methods

We now use the EBI's mapping between individual PDB residues and SwissProt residues to 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`).

Secondly 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.-) where a full EC number has not yet been assigned in the Enzyme database. These mappings are loaded into a second database table.

Queries of the resulting database allow PDB chains to be linked to SwissProt codes and thus to EC numbers by joining the two tables. The complete mappings are then exported from the database as a flat file and in XML. A web query interface is provided at `http://www.bioinf.org.uk/pdbsprotec` allowing queries on the basis of PDB code (optionally with chain identifier), 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 'make' utility, normally used in compiling software to compile only those files that have been updated via time dependencies, 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 (a. the PDB/SwissProt mapping from the EBI; b. the Enzyme database; c. 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.

3 Discussion

PDBSprotEC provides a reliable mapping between PDB chain and EC number in a totally automated fashion. As source data are updated, so the database used by PDBSprotEC 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. In neither case is the mapping downloadable for use in further analysis.

4 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). *Bioinformatics*, **18**, 175–181.

Bairoch, A. (2000). *Nuc. Ac. 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). *Nuc. Ac. 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). *Nuc. Ac. Res.*, **31**, 365–370.

Chamberlin, S. G., Brennan, L., Puddicombe, S. M., Davies, D. E. and Turner, D. L. (2001). *Eur. J. Biochem.*, **268**, 6247–6255.

Etzold, T. and Argos, P. (1993). *Comput. Appl. Biosci.*, **9**, 49–57.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). *Nature (London)*, **358**, 86–89.

Laskowski, R. A. (2001). *Nuc. Ac. 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). *Structure*, **6**, 875–884.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). *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). *Structure*, **5**, 1093–1108.

Pearson, W. R. and Lipman, D. J. (1988). *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). *Proc. Natl. Acad. Sci. USA*, **84**, 8075–8079.