



## 3Dee: a database of protein structural domains

Asim S. Siddiqui<sup>1</sup>, Uwe Dengler<sup>2</sup> and Geoffrey J. Barton<sup>1, 2,\*</sup>

<sup>1</sup>University of Oxford, Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK and <sup>2</sup>EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on May 23, 2000; revised on October 3, 2000; accepted on October 10, 2000

### ABSTRACT

**Summary:** The 3Dee database is a repository of protein structural domains. It stores alternative domain definitions for the same protein, organises domains into sequence and structural hierarchies, contains non-redundant set(s) of sequences and structures, multiple structure alignments for families of domains, and allows previous versions of the database to be regenerated.

**Availability:** 3Dee is accessible on the World Wide Web at the URL <http://barton.ebi.ac.uk/servers/3Dee.html>

**Contact:** [geoff@ebi.ac.uk](mailto:geoff@ebi.ac.uk); <http://barton.ebi.ac.uk>

The Protein Data Bank—PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000) gathers new protein and nucleic acid three-dimensional structures into a common format. In order to make these data useful for many different types of analysis, significant processing is necessary. This processing may add extra information to the data in the form of annotations, but it often includes the definition of binary or multi-way relationships.

The domain is an important intermediate in the protein structural hierarchy. However, domains are not defined systematically in the PDB and sequence and structural relationships between domains are not recorded. The SCOP (Lo Conte *et al.*, 2000), CATH (Orengo *et al.*, 1997), Dali/FSSP (Holm and Sander, 1998) and 3Dee (Dengler *et al.*, 2000) domain databases are approaches to resolve this problem. The unique design considerations for 3Dee were that it should allow alternative domain definitions for the same protein, store multiple structure alignments for families of domains, include derived information and allow previous versions of the database to be regenerated.

Sequence similarity between PDB chains allows them to be clustered into sequence families, but some chains in the same sequence family may have different numbers of domains. Accordingly, sequence families in 3Dee are divided into 'similar domain organisation families', i.e. chains with the same number of equivalent domains.

These chains are split into domain families. Representatives from each domain family are clustered by sequence similarity to give 'domain sequence families', whose representatives are a non-redundant set of domain sequences. Finally, 'domain structure families' cluster these representatives according to the similarity of their three-dimensional structure. Different thresholds of structural similarity give rise to a hierarchy of structurally related domains. Figure 1 shows (a) how basic information stored in a PDB file is retrieved, (b) domain families are created, and (c) 'domain sequence families' and 'domain structure families' are produced.

Each 3Dee domain is tagged by a unique identifier which consists of the PDB code, chain identifier, domain number, definition source and revision number. Alternative domain definitions for the same protein may co-exist.

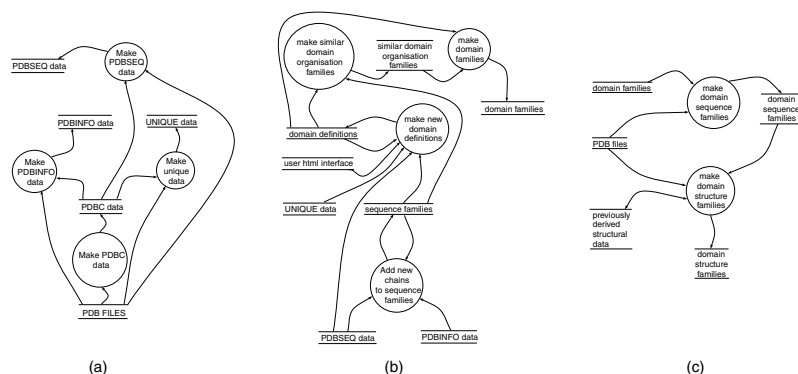
The PDB changes, with approximately 200 new structures deposited each month and a handful removed. Identifying PDB files that are new, modified or to be deleted from the existing version of 3Dee is a straightforward task. However, any changes must be transmitted down all levels of the hierarchy.

In the first stage of a 3Dee database update, data related to old PDB files are stored in backup directories. Domain families are also backed up. Together, this information is sufficient to allow an old version of the database to be regenerated.

When the first release of the database was created, sequence families were derived by complete linkage clustering with OC (Barton, 1997), according to probability scores calculated with the SCANPS sequence comparison program (Barton, 1993). To avoid recalculation for all chains at each update, new chains are either added to existing sequence families or new sequence families are created.

Most of the database creation and update processes (sequence comparison, clustering, defining domains by sequence similarity, structure comparison, etc.) are performed automatically. However, there are no automatic methods that are able to define structural domains accurately and consistently in all proteins. Thus, defining domains had to be carried out, in part, by eye or with

\*To whom correspondence should be addressed. Present address: EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.



**Fig. 1.** Items between two lines are data stores or sources, while those in circles are processes. PDBC data contains information about the chains in a PDB file, PDBSEQ data, the amino acid sequence of the chains, UNIQUE data the residue numbering and PDBINFO data general information (a). The relationships between these data, sequence families and the process of defining domains are illustrated in (b). The dataflow from domain families to ‘domain structure families’ via ‘domain sequence families’ is shown in (c).

reference to the literature. To simplify the domain definition process, http-based client-server software controlling creation, editing and updating of domain definitions in the database was developed. Consistency and error checking programs are incorporated directly into this process to prevent incorrect data from entering the database. Domain definitions may be checked simultaneously by different people in different locations; once domain definitions have been produced, they can be viewed using a RasMol (Sayle and Milner-White, 1995) WWW/PDB interface.

3Dee was first created on the 24th November 1994. Since then, the update techniques have been used to bring the database in line with the PDB seven times; five additional updates were domain definition updates.

The November 1999 release of 3Dee contains 7995 PDB files, 12 458 chains and 18 896 domains, more than three times the original size. A non-redundant set of 1715 domain sequences has been produced which are classified further according to their structural similarity into a conservative 1199 domain structure families.

Pages for each PDB chain available in 3Dee are accessible via a search engine on the World Wide Web. On a chain page, the default domain definition and links to alternative domain definition(s) are provided. If the RasMol interface is installed, it is possible to view the domain(s) defined. One can enter the classification for the default domain definition which offers information about the domain sequence family and links to the structural classification trees. Structural superpositions of domains in the hierarchy can be viewed for clusters above a significant similarity score.

Currently, 3Dee exploits the Unix file system to organise formatted text files. These files are accessed via file-processing applications written in Perl and C. The requirement that 3Dee keeps previous versions of the database accessible, is managed by a simple, custom

revision control system. This implementation has worked effectively for creating and updating of the database. However, the base information in 3Dee is tabular and could benefit from the advantages of a relational database management system.

## ACKNOWLEDGEMENTS

We thank Prof Louise N. Johnson for her encouragement, and Dr Stephen M.J. Searle for helpful discussions. This work was supported by a BBSRC studentship to AS. GJB thanks the European Molecular Biology Laboratory and the Royal Society for support.

## REFERENCES

- Barton, G.J. (1993) An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput. Appl. Biosci.*, **9**, 729–734.
- Barton, G.J. (1997) OC—a cluster analysis program; usage notes.
- 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.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
- Dengler, U., Siddiqui, A.S. and Barton, G.J. (2000) Protein structural domains: analysis of the 3Dee domains database. *Proteins*, in press.
- Holm, L. and Sander, C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
- Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
- 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.
- Sayle, R. and Milner-White, E.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.