492–493 *Nucleic Acids Research, 2003, Vol. 31, No. 1*
DOI: 10.1093/nar/gkg022

PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003

Tamotsu Noguchi* and Yutaka Akiyama

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received September 9, 2002; Accepted September 20, 2002

ABSTRACT

PDB-REPRDB is a database of representative protein chains from the Protein Data Bank (PDB). Started at the Real World Computing Partnership (RWCP) in August 1997, it developed to the present system of PDB-REPRDB. In April 2001, the system was moved to the Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST) (http://www.cbrc.jp/); it is available at http://www.cbrc.jp/pdbreprdb/. The current database includes 33 368 protein chains from 16 682 PDB entries (1 September, 2002), from which are excluded (a) DNA and RNA data, (b) theoretically modeled data, (c) short chains ($l < 40$ residues), or (d) data with non-standard amino acid residues at all residues.

The number of entries including membrane protein structures in the PDB has increased rapidly with determination of numbers of membrane protein structures because of improved X-ray crystallography, NMR, and electron microscopic experimental techniques. Since many protein structure studies must address globular and membrane proteins separately, this new elimination factor, which excludes membrane protein chains, is introduced in the PDB-REPRDB system. Moreover, the PDB-REPRDB system for membrane protein chains begins at the same URL. The current membrane database includes 551 protein chains, including membrane domains in the SCOP database of release 1.59 (15 May, 2002).

INTRODUCTION

Protein structure data in PDB (1) are used actively in studies of protein function, evolution and structure prediction, but not all data are competent for the purpose of protein structure analysis. The number of entries in the PDB has increased rapidly due to international structural genomic initiatives. Many entries have insufficiently-refined coordinate data, perhaps due to insufficient resolution in the X-ray crystallography, NMR spectroscopy or electron microscopy. In many cases, such imperfect data should be eliminated *ex ante* to achieve an accurate result. Moreover, large numbers of protein chains in PDB are similar in terms of sequence or structural similarity. For an unbiased analysis, one may have to classify these chains and select only one representative from each group of similar chains.

On the other hand, local structural diversity is informative to investigate principles of local conformation of proteins. Several protein chains whose similarity is greater than 90% have been found to exhibit structural diversities that are particularly local. Local structural diversities have been found at insertion, deletion, or mutation sites since these sequence modifications cause structural changes. Moreover, a protein would change local conformation by docking with other molecules (e.g., DNA, RNA or another protein) or heterozygous atoms.

PDB-REPRDB (2,3) is a database of representative protein chains from the PDB, whose criteria used to select the representatives are: (a) quality of atomic coordinate data, (b) sequence uniqueness, and (c) conformation uniqueness that is particularly local. We introduced the sequence identity (ID%) and the maximum distance between superimposed pairs of atoms from the two structures (Dmax) as respective measures of sequence and structural similarities, which is more sensitive to detection of local structural diversity than root mean square deviation (RMSD). PDB-REPRDB was initially started at the Real World Computing Partnership (RWCP) in August 1997; it subsequently developed to the present system of PDB-REPRDB, which assures a quick selection of representative chains sets based on users' requirements by an interactive system using WWW. In April 2001, the system was moved to the Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST) (http://www.cbrc.jp/); it is available at the new WWW server (http://www.cbrc.jp/pdbreprdb/).

We have now improved the PDB-REPRDB system for membrane protein chains and developed the new PDB-REPRDB system for membrane protein chains.

RECENT DEVELOPMENTS

In recent years, the number of entries including membrane protein structures in the PDB has increased rapidly with the

*To whom correspondence should be addressed. Tel: +81 3 3599 8041; Fax: +81 3 3599 8081; Email: noguchi-tamotsu@aist.go.jp determination of numbers of membrane protein structures because of improved X-ray crystallography, NMR, and electron microscopic experimental techniques. Since protein structure researchers deal separately with globular proteins and membrane proteins, membrane proteins should be separate from globular proteins.

One recent improvement is the introduction of a new factor of elimination, which does not include membrane protein chains, for selecting representatives on the top page of the present PDB-REPRDB. The membrane protein chain, including a domain defined by the SCOP (4,5), is eliminated by this factor.

Moreover, the PDB-REPRDB for membrane protein chains, which selects representatives from membrane protein chains including a membrane domain defined by the SCOP, has been developed for researchers for membrane proteins.

CONTENT OF DATABASE

The current database of PDB-REPRDB includes 33 368 protein chains from 16 682 PDB entries (1 September, 2002), from which are excluded (a) DNA and RNA data, (b) theoretically modeled data, (c) short chains ($l < 40$ residues), or (d) data with non-standard amino acid residues at all residues. The number of representative chains, which were selected on several pairs of sequence and structural similarity parameters, is shown in a table at the sample page of PDB-REPRDB. Moreover, tables of older releases are linked to the page.

The current database for the membrane includes 551 protein chains, which include membrane domains in SCOP database of release 1.59 (15 May, 2002).

USAGE OF PDB-REPRDB

The present system of PDB-REPRDB is designed so that the user may obtain a quick selection of representative chains from PDB. The WWW interface provides enhanced freedom in setting parameters, such as cut-off scores of sequence and structural similarity. Users can eliminate unnecessary chains from the PDB chain list by setting threshold values; users can also change the priority of nine factors (resolution, R-factor, number of chain breaks, ratio of non-standard amino acid residues, ratio of residues with only Cα coordinates, ratio of residues with only backbone coordinates, number of residues, whether mutant or wild and whether complex or not). Moreover, users can select whether or not to include entries by NMR experimental techniques and whether or not to include membrane proteins by setting a flag of NMR and membrane, respectively. The membrane flag is not included in the system of PDB-REPRDB for membrane protein chains. One can obtain a representative list and classification data for protein chains from the systems.

The representative list includes the factor information mentioned above, EC number and compound in PDB. 'ID' sections are hyperlinked to data on classified groups; also, a graphic representation of the three-dimensional structure can be displayed using the RasMol program by clicking on '*'. Furthermore, 'EC number' sections are hyperlinked to the ENZYME section of LIGAND: database of chemical compounds and reactions in biological pathways (http://www.genome.ad.jp/ligand) (6,7).

FUTURE WORK

We continue to expand databases for PDB-REPRDB according to the increase of PDB and the update of SCOP. The PDB-REPRDB for all chains will be updated approximately once every two months.

We plan to indicate the SCOP concise classification strings (sccs) to the list of PDB-REPRDB to show the relation between SCOP and PDB-REPRDB. The user will be able to find local structural diversity between proteins of the same family by the sccs.

ACKNOWLEDGEMENTS

We thank Dr Susumu Goto and Prof. Minoru Kanehisa at the Institute for Chemical Research, Kyoto University, for their support.

REFERENCES

1. 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.
2. Noguchi,T., Onizuka,K., Ando,M., Matsuda,H. and Akiyama,Y. (2000) Quick selection of representative protein chain sets based on customizable requirements. *Bioinformatics*, **16**, 520–526.
3. Noguchi,T., Matsuda,H. and Akiyama,Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res.*, **29**, 219–220.
4. 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.
5. Conte,L.L., Brenner,S.E., Hubbard,T., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
6. Suyama,M., Ogiwara,A., Nishioka,T. and Oda,J. (1993) Searching for amino acid sequence motifs among enzymes: the Enzyme-Reaction Database. *Comput. Appl. Biosci.*, **9**, 9–15.
7. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compound and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.