Nucleic Acids Research, 2001, Vol. 29, No. 1 **219–220**

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

Tamotsu Noguchi[1,2,*], Hideo Matsuda[2] and Yutaka Akiyama[1]

[1]Computational Biology Research Laboratory, Electrotechnical Laboratory, 1-1-4 Umezono, Tsukuba, Ibaraki 305-5868, Japan and [2]Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

Received September 1, 2000; Revised and Accepted October 4, 2000

ABSTRACT

PDB-REPRDB is a database of representative protein chains from the Protein Data Bank (PDB). The previous version of PDB-REPRDB provided 48 representative sets, whose similarity criteria were predetermined, on the WWW. The current version is designed so that the user may obtain a quick selection of representative chains from PDB. The selection of representative chains can be dynamically configured according to the user's requirement. The WWW interface provides a large degree of freedom in setting parameters, such as cut-off scores of sequence and structural similarity. One can obtain a representative list and classification data of protein chains from the system. The current database includes 20 457 protein chains from PDB entries (August 6, 2000). The system for PDB-REPRDB is available at the Parallel Protein Information Analysis system (PAPIA) WWW server (http://www.rwcp.or.jp/papia/).

INTRODUCTION

The protein structure data in PDB (1) are being used actively in studies of protein function, evolution and structure prediction, but not all the data are competent for the purpose of protein structure analysis. A lot of entries have insufficiently-refined coordinate data, perhaps due to insufficient resolution in the X-ray crystallography or NMR spectroscopy. In many cases one may want to eliminate the imperfect data beforehand to achieve an accurate result. Moreover, a great deal 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.

At present, several classification databases (2–7) have been proposed and are available on the WWW, but the selected set would not reflect local structural diversities between members of a protein family. Local structural diversity is informative to investigate the principles of the local conformation of proteins. Local structural diversities have also been found at insertion, deletion or mutation sites, since these sequence modifications cause structural changes.

We earlier reported 'PDB-REPRDB', a database of representative protein chains selected from PDB (8). The criteria used to select the representatives were: (i) quality of atomic coordinate data, (ii) sequence uniqueness and (iii) 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 the respective measures of sequence and structural similarities, which is more sensitive to the detection of the local structural diversity than root mean square deviation (RMSD).

The previous version of PDB-REPRDB provided 48 representative sets (eight criteria for sequence similarity: ID% ≥ 25–95% with 10% increments and six criteria for structural similarity: Dmax ≤ 10–50 Å with 10 Å increments and ∞: differences in structure not considered) on the WWW. However, the sets were insufficient in number to satisfy users researching protein structures by various methods.

The current version of PDB-REPRDB assures a quick selection of representative chains sets based on the user's requirement by the interactive system using a WWW user interface (9).

METHOD

We define the similarities between protein chains by means of ID%, RMSD and Dmax. These similarity values are calculated for each pair of protein chains. First, a pair of chains is aligned by the pairwise sequence alignment developed by Needleman and Wunsch (10) and ID% is calculated from the result of alignment. Next, each pairs of Cα atoms in the aligned residues are superimposed by the least square fitting procedure (11), and RMSD and Dmax are calculated from the superposition. This procedure is executed beforehand every time a new PDB is released, and the interactive system classifies those chains and selects the representatives using the similarity data.

CURRENT DATABASE

The system for PDB-REPRDB is available at the PAPIA WWW server (http://www.rwcp.or.jp/papia/) (12). The PDB-REPRDB is currently selected from 20 457 chains, which do not include (i) DNA and RNA data, (ii) theoretically modeled data, (iii) short chains (l < 40 residues) or (iv) data with non-standard amino acid residues at all residues. The user can elim-

*To whom correspondence should be addressed. Tel: +81 298 61 5080; Fax: +81 298 61 5722; Email: tnoguchi@etl.go.jp

inate unnecessary chains from the PDB chain list by setting threshold values, and change the priority of factors (Table 1) for selecting representatives on the top page.

Table 1. Elimination factors and the default priority

Factors for elimination	Default priority
Resolution	1
R-factor	2
Number of chain breaks	3
Ratio of non-standard amino acid residues	4
Ratio of residues with only Cα coordinates	5
Ratio of residues with only backbone coordinates	6
Number of residues	7
Include mutant	8
Include complex	9
Include NMR	–

A sequence similarity parameter or pairs of sequence and structural similarity parameters (e.g. ID% ≥ 30% and RMSD ≤ 15 Å, ID% ≥ 90% and Dmax ≤ 5 Å) are selected and set the values on the following page. As the result, a list of representative chains and the classification data of chains for the parameters can be obtained from the system. The numbers of representative chains, which selected on several pairs of sequence and structural similarity parameters, are shown in Table 2.

Table 2. Number of representative chains

Sequence identity (ID%)	Number of chains Dmax (Å)			
	≥10	≥30	≥50	∞
≤30	3112	2177	2108	2098
≤50	3479	2940	2898	2888
≤70	3941	3480	3442	3434
≤90	4674	4295	4262	4257

'ID' (PDB entry ID + chain ID) sections on the list of representative chains are hyperlinked to the screen, which contains data on the classified groups and a graphic representation of the three-dimensional structure can be displayed using the RasMol program, by clicking on '*'. Furthermore, 'ECnumber' sections are hyperlinked to LIGAND (Ligand chemical database for enzyme reactions) (13), which is one of the databases supported by DBGET/LinkDB (14) on GenomeNet in Japan. The classification data are presented on one page, in which each representative chain and the similar chains in its group are described by 'ID' on a single line. Each 'ID' is hyperlinked with the PDB on the DBGET/LinkDB; clicking it will show the contents of the corresponding PDB entry.

ACKNOWLEDGEMENTS

We thank Dr Susumu Goto and Prof. Minoru Kanehisa at Institute for Chemical Research, Kyoto University for their support. The computation environment is provided by the Tsukuba Research Center, Real World Computing Partnership.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, 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.
2. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
3. Sander,C. and Schneider,R. (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
4. Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600–3609.
5. 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.
6. 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.
7. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
8. Noguchi,T., Onizuka,K., Akiyama,Y. and Saito,M. (1997) PDB-REPRDB: A database of representative protein chains in PDB (Protein Data Bank). *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 214–217.
9. 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.
10. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
11. Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **A34**, 827–828.
12. Akiyama,Y., Onizuka,K., Noguchi,T. and Ando,M. (1998) Parallel Protein Information Analysis (PAPIA) system running on a 64-node PC cluster. *Proceedings of the Ninth Workshop on Genome Informatics*, Universal Academy Press, pp. 131–140.
13. Goto,S., Nishioka,T. and Kanehisa,M. (1999) LIGAND database for enzymes, compounds and reactions. *Nucleic Acids Res.*, **27**, 377–379.
14. Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) DBGET/LinkDB: An integrated database retrieval system. *Pac. Symp. Biocomput.*, **1998**, 683–694.