PDB-REPRDB: A Database of Representative Protein Chains in PDB (Protein Data Bank)

Tamotsu NOGUCHI, Kentaro ONIZUKA, Yutaka AKIYAMA, and Minoru SAITO

Parallel Application Laboratory, Tsukuba Research Center
Real World Computing Partnership
Tsukuba Mitsui Bldg., 1-6-1 Takezono, Tsukuba 305, Japan
E-mail: noguchi@trc.rwcp.or.jp

Abstract

We propose a novel set of 'representative' protein chains in PDB, where not only sequential but also structural similarities are taken into account.
Hobohm *et al.* have already proposed "PDB_SELECT", which eliminates redundant chains based solely on sequence similarity. "PDB_SELECT" is frequently updated and the latest version is available at EMBL WWW server.
In our set of entries "PDB-REPRDB," however, structural similarities are also considered, in order not to overlook local conformation diversity within a group of sequentially similar chains. Our set guarantees that every representative is the best among that similar protein group, regarding experimental or structure-determination quality (*i.e.* resolution and R-value).
The first version (based on PDB Release 70) of PDB-REPRDB was released in 1995 and the second version (PDB Release 78) will be available by April 1997.

Keywords: PDB (Protein Data Bank); representative protein chains; sequential and structural similarity

Introduction

Necessity of Excluding Inappropriate Entries

The Protein Data Bank (PDB) (Bernstein *et al.* 1997) is a rich library of atomic-coordinate data of biological macromolecules. That has been an indispensable resource for computational study of protein structure.

The PDB entries has been increasing rapidly, though not all entries are competent for the purpose of computational protein-structure analysis. A lot of entries have insufficiently-refined coordinate data perhaps with insufficient resolution in X-ray crystallography. In many cases we should better eliminate those imperfect data beforehand for an accurate and unbiased analysis.

Considering that a protein in solvent usually has a different conformation from that in the crystal, such PDB entries, particularly of NMR, should be distinguished. If not, the plural structure models determined by NMR cause another practical problem, that is "Which should we select out of many models?"

Listing Representative Chains

A great deal of entries in PDB has some or many similar entries in terms of structural or sequential similarity (*e.g.* lysozyme together with its mutants occupy more than 300 entries).

To avoid a result strongly biased toward populous chain-families in statistical analyses, we should consider the population of each similar chain group. Some weighting scheme might be theoretically preferable, though it is none-the-less difficult in practice.

The simplest strategy is to select a 'representative' entry out of each similar chain group. The resultant set of representatives, here, non-redundantly cover the whole spectrum of PDB. The criteria of the selection should be different according to the purpose of the analysis.

Hobohm *et al.* have proposed a representative set of protein chains (Hobohm *et al.* 1992). The set, "PDB_SELECT" which has been being updated and open to public on the anonymous ftp site (Hobohm & Sander 1994; Hobohm & Sander), is widely regarded as a current standard in the community.

PDB_SELECT's strategy of selection is, however, solely based on the sequential similarity. The sequentially similar chains are, thus, automatically assumed to have the similar structure. As the result, only one representative is selected out of sequentially similar group regardless of structural diversity.

This strategy seems rational in terms of equality in homologous-sequence elimination, and easy implementation, though chances are that the selected set would overlook diversity of local structures in a phylogenetically related protein family. Local structure diversity is non-the-less informative to investigate the formation principles of protein local conformation. (Fig. 1).

We consider these local-conformation-diversity to be conserved in the representative set.

On the other hand, several groups have constructed

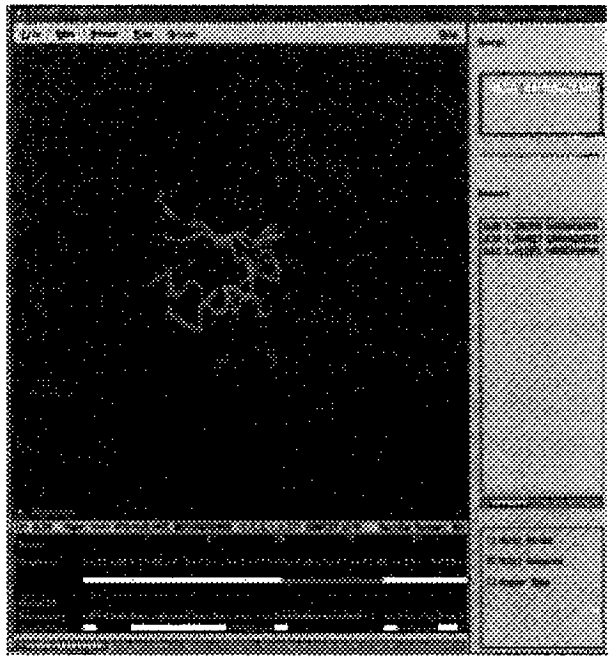

Figure 1: Example of local conformation disagreements: In this case, a query sequence, Lactate Dehydrogenase (8LDH), is superimposed to three related chains. The sequence of 1LDM is 100% identical to 8LDH, while a turn structure (residues 101-106) shown in the upper-right region disagrees with that part of 8LDH, due to the different complex formation. The other two (2LDX and 1LLD) with less sequence identity better fit to 8LDH.

useful databases: *e.g.* CATH proposed by Orengo *et al.* and SCOP by Murzin *et al.* These classified the protein chains regarding both sequential and structural similarity. CATH and SCOP are available on WWW sites (Orengo *et al.*; Murzin *et al.* 1995) and enthusiastically utilized by structure biologists in the world. What they did is, however, actually hierarchical classification of protein chains. They did not selected representatives for the purposes of statistically unbiased structure analysis.

PDB-REPRDB: A new representative chain database

In this paper, we report our "PDB-REPRDB," a new database of representative protein chains selected from PDB. The criteria of selecting the representatives are, a) quality of atomic coordinate data, b) sequence uniqueness, and c) conformation uniqueness particularly local.

The first version of PDB-REPRDB consists of 763 representative chains from PDB Release 70 (Oct. 1994) and was released in July 1995 on GenomeNet WWW server (Noguchi 1995). Each entry has hyperlinks to that entry in PDB.

The new version of PDB-REPRDB selected from PDB Release 78 (Oct. 1996) will be scheduled to be released (Noguchi *et al.*). The selection policy remains the same, while the selection procedure is almost completely automated by sophisticated algorithms. The system parallelization is one thing yet to do as a future work.

Methods

The new version of PDB-REPRDB is derived from PDB Release 78 (Oct. 1996). The procedure of selecting representative protein chain is as follows.

Phase 1: Excluding inappropriate entries

Those PDB entries which matches any of the conditions below are excluded in this phase.

a) DNA and RNA entries,

b) data derived by NMR spectroscopy,

c) theoretically modeled data,

d) short chains ($l < 40$ residues),

e) data without backbone coordinates at all residues,

f) data without side chain coordinates at all residues, or

g) data without refinement (by X-PLOR, TNT, etc.).

The current scope of our database is concentrated on those protein structures determined by X-ray crystallography. Hence, those entries of (a) DNA and RNA, or determined by (b) NMR, or (c) theoretical modeling are excluded.

Phase 2: Sorting chains with respect to data quality

All chains are extracted from each entry selected through phase 1. These chains are subject to the selection in phase 2. This phase actually sort the chains according to the data quality.

First, the selected chains are classified into two classes. Class A chains are those with good resolution ($\leq$ 3.0 Å) and good R-Factor ($\leq$ 0.3). Other chains are classified into class B.

Second, we sort the chains with respect to the resolution of structure determination within each class (A and B). The chains with the same resolution are further sorted by R-Factor value. When plural chains have the same resolution and R-Factor, those are sorted by:

1. the number of chain breaks (the less the better),
2. the number of non-standard amino acid residues (the less the better),
3. the number of residues without backbone coordinates (the less the better),
4. the number of residues without side chain coordinates (the less the better),
5. whether mutant or wild (the wild type has priority),
6. whether complex or not (the non-complex has priority), and
7. alphabetical order of the entry name. (*e.g.* 1MCD < 1MCE, 5AT1A < 5AT1C)

Phase 3: Elimination of the redundant chains

In this elimination phase, those chains with better quality have the priority to be the representatives. The first chain in the sorted list is to be the first representative because the chain has the best quality. The algorithm to determine all representatives is simple.

Suppose we have already selected N representatives from the sorted list. Now that the sorted list does not contain the chains 1) which have already been selected and taken out as the representatives, 2) which have already been eliminated through the selection procedure of the N representatives.

Thus, the first chain remained in the sorted list has the highest priority to be the next representative. We check the "similarity" between the first chain of the list and each of the already selected representatives. If the first chain is not similar to any of the selected representatives, the first chain becomes the $(N+1)$-st representative. If the first chain is similar to at least one of the representatives, the chain is eliminated, and then, the second chain comes to the first of the sorted list. This procedure repeats until the sorted list goes to null.

The representatives are firstly selected from class A chains. Class B chains are also sorted and appended to the tail of the list, thus these chains are chosen only if there are no similar chain available in class A.

In the similarity check, we consider the sequence is NOT similar,

- if the sequential identity is less than a certain threshold value ($\leq$ 90%, 80%, 75%, *etc.*), where the sequence identity is measured by FASTA algorithm (Pearson & Lipman 1988; Pearson 1990),
- or, if the maximum distance between the superimposed pair of atoms each from the two structures that we are looking at is greater t han a certain threshold ($\geq$ 10 Å in this paper, 9 Å , *etc.*).

Before superimposing the two structures, we align the two sequences by the pairwise sequence alignment. The matched sites in the alignment are superimposed by the least square fitting procedure (Kabsch 1976; Kabsch 1978).

Finally, all the chains (both in class A and B) are classified into protein-chain groups, where each chain is classified into the group whose representative chain is sequentially nearest to the chain.

Table 1: Comparison of PDB-REPRDB and PDB_SELECT based on PDB Release 78

	Number of chains	
	PDB-REPRDB ver.2.0 (1997)	PDB_SELECT (1997)
Total	1089	1405
X-Ray Data	1089	1184
NMR Data	0	214
Other Data	0	7

Note: Threshold sequence identity (ID%) here is 75%.

Table 2: Threshold sequence identity (ID%) and number of selected chains

Threshold sequence identity (ID%)	Number of chains	
	PDB-REPRDB ver.2.0 (1997)	PDB_SELECT (1997)
75	1089	1405
80	1139	-
85	-	1546
90	1251	-
95	-	1718

Note: Both databases are based on PDB Release 78.

Comparison with PDB_SELECT

To clarify the difference between our representative chain set and PDB_SELECT (Hobohm *et al.* 1992; Hobohm & Sander 1994; Hobohm & Sander), we compared the number of selected chains of ours and PDB_SELECT as shown in Table 1. Both representative sets are based on PDB release 78.

When the threshold sequence-identity (ID%) for defining a sequentially similar group is 75%, the number of representative chains is 1080 in our PDB-REPRDB where the structures determined by NMR and MODEL were eliminated. The number of chains in PDB_SELECT is 1405 in total, and 1184 for X-ray but it includes many low quality data.

Table 2 shows the number of selected chains varied with the threshold value for ID%.

Discussion

The first version of PDB-REPRDB already had such chains that are sequentially similar to other chains but have different local structure. The second version has even more chains in this respect.

Considering the increasing structures determined by NMR spectroscopy in PDB, it is not a good strategy to ever eliminate those NMR structures from the representative set. Since our representative set considers structural similarity, it is required to establish a method to compute the similarity between structures by X-Ray diffraction and NMR. Then we will have greater number of structurally similar groups than the current dataset.

In addition, we need even more automated selection method of representative chains from the large number of entries in PDB. We are going to implement a fully parallelized system which assures a quick selection of representative chains on user's demand, or as soon as the new version of PDB is released.

Conclusion

In this paper, we proposed PDB-REPRDB (a novel 'representative' database of protein-chains from PDB), whose selective criterion is based on sequential and structural difference.

It was first for our own research activities (*i.e.* protein secondary structure prediction, local structure classification) that we developed the first version of PDB-REPRDB. We knew, then, that the selected dataset was strongly recommendable to the researchers of protein-structure fields. The point is that our dataset discriminates sequentially similar chains with meaningful local structure diversity (due to insertions, deletions, or mutations), and even discriminates small structural change caused by complex formation.

The new version about to be released will contribute toward improvements of the quality of protein structure analyses. We hope that our PDB-REPRDB will be used as widely as other pioneering database system like PDB_SELECT, SCOP and CATH.

Acknowledgments

We thank Dr. Susumu Goto and Prof. Minoru Kanehisa at Institute for Chemical Research, Kyoto University for useful discussions and suggestions.

References

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. *Journal of Molecular Biology* **112:** 535-542. `http://www.pdb.bnl.gov/`

Hobohm, U.; Scharf, M.; Schneider, R.; and Sander, C. 1992. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Science* **1:** 409-417.

Hobohm, U.; and Sander, C. 1994.
Enlarged representative set of protein structures. *Protein Science* **3:** 522.

Hobohm, U.; and Sander, C. PDB_SELECT: Representative list of PDB chain identifiers.
`http://www.sander.embl-heidelberg.de/pdbsel/explanation.html`

Orengo, C. A.; Michie, A. D.; Jones, S.; Swindells, M. B.; Jones, D. T.; and Thornton, J. M.
CATH: Protein Structure Classification, version 1.0.
`http://www.biochem.ucl.ac.uk/bsm/cath`

Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. 1995. scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **247:** 536-540.
`http://scop.mrc-lmb.cam.ac.uk/scop/`

Noguchi, T. 1995. PDB-REPRDB ver.1.0 served on GenomeNet WWW Server.
`http://www.genome.ad.jp/htbin/show_pdbreprdb`

Noguchi, T.; Onizuka, K.; Akiyama, Y.; and Saito, M. 1997. PDB-REPRDB ver.2.0 served on RWCP WWW Server. `http://www.rwcp.or.jp/lab/mpap/`

Pearson, W. R.; Lipman, D. J. 1988. Improved Tools for Biological Sequence Analysis. *Proc. Natl. Acad. Sci.* **85:** 2444-2448.

Pearson, W. R. 1990. Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Methods in Enzymology* **183:** 63-98.

Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **A32:** 922-923.

Kabsch, W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **A34:** 827-828.