Protein Science (1994), *3*:522–524. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Enlarged representative set of protein structures

UWE HOBOHM AND CHRIS SANDER
European Molecular Biology Laboratory, 69012 Heidelberg, Germany

(RECEIVED September 16, 1993; ACCEPTED December 23, 1993)

Abstract: To reduce redundancy in the Protein Data Bank of 3D protein structures, which is caused by many homologous proteins in the data bank, we have selected a representative set of structures. The selection algorithm was designed to (1) select as many nonhomologous structures as possible, and (2) to select structures of good quality. The representative set may reduce time and effort in statistical analyses.

Keywords: NMR; PDB; representative protein data set; X-ray crystallography.

Representative selection of proteins of known 3-dimensional structure

There is considerable redundancy in the Protein Data Bank (PDB) (Bernstein et al., 1977) of 3-dimensional structures. For example, there currently are atomic coordinates for about 77 globins, 61 immunoglobulins, and 9 structures of phage T4 lysozyme, including many engineered mutants. Although the wealth of data bears witness to the progress achieved by protein crystallographers and NMR spectroscopists, an overview of the spectrum of known protein structures and certain statistical analyses of protein structures require nonredundant data. To meet this need, we have developed algorithms to select from PDB (or from sequence databases) representative subsets that aim to minimize redundancy and maximize coverage (Hobohm et al., 1992). The result, in the form of a list of PDB identifiers, was published about $1\frac{1}{2}$ years ago and has been very useful, e.g., in developing better algorithms for secondary structure prediction (Rost & Sander, 1993). Since then, there has been rapid growth of the number of known protein structures and, in addition, a sudden surge of preliminary data sets released by the PDB—from about 600 in early 1992, there are now about 2,000 PDB coordinate data sets, a 3-fold increase in 18 months. It was therefore time to update the representative list and, based on experience gained since the original publication, to refine the criteria for selection. The result is an increase from 155 to 301 (95%) in the number of sequence-unique proteins. The current "25%-list" (based on PDB release, December 1993), i.e., a list of protein chains with less than 25% sequence identity, was constrained to be downward compatible with the "30% list" published in 1992 (Hobohm et al., 1992). Some adjustments were a result of the more stringent threshold and new quality criteria, e.g., replacement of a data set by one with identical sequence but higher resolution. Downward compatibility will reduce the number of changes in user applications upon new releases of the list and will be maintained in the future.

The list is provided in Table 1 and on the Diskette Appendix and is also available on the EMBL file transfer (ftp) and e-mail servers.

Quality control of selected data sets

Given a choice between 2 different data sets, from which only 1 can be selected into the list, one would like to use the one of higher quality. For instance, one wishes to avoid the use of data sets based on very preliminary or incomplete data and restrict the selection to proteins that are not small and consist primarily of the standard 20 amino acids. These quality goals are implemented at 3 levels, an initial filter, a "softexclude" flag for marginal data, and a quality index. The quality index Q of a protein chain is calculated as

$$Q = r + R/20,$$

where r is resolution (Å) and R is R-factor (percent). Of 2 chains, the one with higher Q is considered to be of lower quality.

As an initial filter, we eliminate all chains with: (1) 100% sequence identity on the entire length to another chain, but lower quality, (2) more than 5% of nonstandard or unknown amino acids (typically termed "UNK" in the PDB files), (3) a length of less than 30 residues, (4) a resolution worse than 3.5 Å, (5) an R-factor of more than 30%, and (6) models not based directly on X-ray or NMR data, e.g., models built by homology. The values for cutoffs were chosen by experience and can be changed to meet special requirements.

Second, some chains are flagged for preferred exclusion in the subsequent selection procedure ("softexclude flag"). Such chains remain in the list only if no homologous chain exists. These are chains for which (1) the number of residues with sidechain co-

Reprint requests to: Uwe Hobohm, European Molecular Biology Laboratory, 69012 Heidelberg, Germany; e-mail: hobohm@embl-heidelberg.de.

Table 1. *Representative list of Protein Data Bank chain identifiers*[a]

1AAF	1AAIB	1AAK	1AAPA	1ABA	1ABH	1ABK	1ABMA	1ACE	1ADS	1AOZA	1APS
1ARB	1ATNA	1ATX	1AVHA	1AYH	1BAA	1BABB	1BARB	1BBA	1BBHA	1BBL	1BBO
1BBPA	1BBT1	1BBT2	1BGC	1BOP	1BOVA	1BRD	1BTC	1BW4	1C2RA	1C5A	1CAJ
1CAUA	1CAUB	1CBN	1CBP	1CD8	1CDTA	1CID	1CIS	1CMBA	1COBA	1COX	1CPCA
1CPCL	1CSEI	1CTAA	1CTC	1D66A	1DFNA	1DHR	1DNKA	1DPI	1DRI	1EAF	1ECO
1EFM	1EGF	1END	1EPS	1ERP	1EZM	1FAS	1FBAA	1FC1A	1FC2C	1FCS	1FDD
1FHA	1FIAB	1FNR	1FXIA	1GKY	1GLAF	1GLAG	1GLT	1GMFA	1GMPA	1GOX	1GPB
1GPS	1GRCA	1GRDA	1GSGP	1GSSA	1HC1	1HCC	1HDDC	1HGEB	1HIGA	1HILA	1HIVA
1HLEB	1HSBA	1HSDA	1IFA	1IFC	1ISUA	1IXA	1L92	1LAP	1LE4	1LGAA	1~LIG
1LTSA	1LTSC	1LTSD	1MAMH	1MDAA	1MDAH	1MDC	1MHU	1MINA	1MINB	1MONA	1MRRA
1MS2A	1MUP	1NIPB	1NRCA	1NRD	1NXB	1OFV	1OMF	1OMP	1OSA	1OVB	1PAFA
1PAZ	1PCDA	1PDA	1PDC	1PDE	1PFKA	1PHB	1PHH	1PHY	1POA	1POC	1PPBA
1PPFE	1PPN	1PRCC	1PRCM	1PRF	1PTE	1PYP	1R094	1R1A2	1R1EE	1RBP	1RCB
1REA	1RHD	1RND	1RPRA	1RRO	1RVEA	1S01	1SAS	1SBP	1SGT	1SHAA	1SHFA
1SMRA	1SNC	1TABI	1TEN	1TFG	1TFI	1TGL	1TGSI	1THO	1TIE	1TLK	1TMD
1TNFA	1TPT	1TRB	1TREA	1TROA	1TTBA	1ULA	1UTG	1VAAB	1VSGA	1WSYA	1WSYB
1XIMA	1YCC	256BA	2AAA	2ACHA	2ACHB	2ATCB	2AVIA	2AZAA	2BDS	2BPA1	2BPA2
2BPA3	2CBH	2CCYA	2CDV	2CPL	2CRD	2CRO	2CTS	2CYP	2DPV	2ECH	2ER7E
2GB1	2GLSA	2HAD	2HHMA	2HHRC	2HIPA	2HPDA	2IHL	2ILA	2LALA	2LALB	2LIV
2MADL	2MEV1	2MEV4	2MHR	2MHU	2MNR	2MSBA	2PF2	2PIA	2PLV1	2PLV3	2PMGA
2POR	2RN2	2SCPA	2SGA	2SN3	2SNV	2STV	2TBVA	2TMVP	2ZTAA	3ADK	3B5C
3CBH	3CD4A	3CHY	3CLA	3DFR	3DPA	3GAPA	3GBP	3GRS	3IL8	3PGK	3PGM
3RUBS	3SC2A	3SC2B	3SGBI	3SICI	4BLMA	4CPAI	4ENL	4FXN	4GCR	4GPD1	4HTCI
4ICD	4INSB	4RCRH	4SBVA	4SGBI	4TGF	4TMS	4TS1A	5FBPA	5NN9	5P21	7AATA
7APIB	7ZNF	8ABP	8ACN	8ADH	8ATCA	8CATA	8I1B	8RXNA	9LDTA	9RNT	9RUBB
9WGAA											

[a] No pair of protein chains in the list has more than 25% identical residues after sequence alignment, using a standard Smith and Waterman (1981) alignment algorithm with gap open penalty 3.0 and gap elongation penalty 0.05. The Diskette Appendix contains this listing, along with additional data about the proteins (e.g., number of residues, crystallographic resolution, crystallographic *R*-factor).

ordinates is less than 90% of the sequence length (e.g., backbone-only structures); (2) the number of residues with backbone coordinates is less than 90% of the sequence length (e.g., C-α-only structures); (3) the number of alanine plus glycine residues is higher than 40% of the sequence length, e.g., structures with unknown sequence modeled as polyalanine; and (4) no data for resolution or *R*-factor are available.

Selection procedure

The new representative list of PDB chain identifiers was produced using "algorithm 2" of Hobohm et al. (1992). This algorithm removes redundant protein chains 1 by 1, following a strategy called "greedy" by computer scientists: the chain with the largest number of neighbors is removed, until no neighbors are left. Neighbors are here defined as pairs of chains with a sequence identity above 25% (or for alignments shorter than 80 residues, with a sequence identity above the significance threshold derived by Sander & Schneider [1991]).

During the run of the selection procedure an ambiguity left by the algorithm may occur when more than 1 protein has the same number of neighbors. At this stage, quality control and downward compatibility are achieved by preferably removing the data set of lower quality and by preferably keeping chains that have been in a previous list.

To select from a group of chains with the same number of neighbors ("exclusion candidates"), a hierarchy of selection criteria was applied: (1) First, remove a chain with a "softexclude" flag. If there is no such chain, (2) remove the chain with lowest quality. If there is more than 1 chain with the lowest quality, (3) remove a chain that has no "softinclude" flag (chains from a previous list are flagged with a "softinclude" flag). If there is more than 1 chain left, (4) remove a chain with high PDB-identifier (i.e., prefer 6TIM over 2TIM). If there is more than 1 chain left, (5) remove a chain with alphabetically higher chain identifier (i.e., prefer 3HLA-A over 3HLA-B).

For special requirements it may be necessary to have chains of individual choice in the list. To meet such a requirement, a subset of chains can be flagged with a "hardinclude" flag. These chains are not removed during the selection procedure, and relations between "hardincluded" chains are not considered. The resulting list will contain all "hardinclude"-flagged chains and as many other chains as possible that do not violate the sequence identity criterion. Special purpose lists can be obtained on request from the authors.

List of homologous chains

For each chain in the "25%-list," we list sequence-related chains at the bottom of the data file (on the Diskette Appendix and on the EMBL file server), i.e., those chains that belong to the same structural family as determined by the HSSP significance threshold (Sander & Schneider, 1991). Because the 25% threshold is on the conservative side, these families may include some false positives, i.e., families with a sequence identity between 25–30% may or may not have the same structure.

Lists at stringency levels from 25% to 90%

A representative selection of proteins in which no 2 proteins have more than 25% sequence identity may be unnecessarily stringent for some purposes. One may want to allow some redundancy of information, e.g., up to 50% identical residues in any pair (Kabsch & Sander, 1983). We have therefore used the same algorithm at gradually decreasing levels of stringency, i.e., allowing pairs with 30%, 35%, 40%, and so on up to 90% identical residues. The lists are incremental: a list with a less stringent similarity threshold includes all proteins in any of the lists with more stringent thresholds. This was done so that a user of the selection can reuse calculations done at a more stringent level (smaller selection) when going to a lower level of stringency (larger selection).

The resulting selection is substantially larger than $1\frac{1}{2}$ years ago: whereas in early 1992 the "30%-list" had 155 proteins (of 600), the new "30%-list" contains 338 proteins (of about 2,000). The new "25%-list" contains 301 protein chains. The redundancy of information in the PDB, as measured by the relative size of the selection, has increased from 4 to 1 to about 7 to 1. In other words, the number of representative structures increases considerably more slowly than the overall number of known structures. Because the representative list is now about a factor of 7 smaller than the entire database, counted as the number of chains, much time and effort may be saved in restricting analyses or statistical evaluations to the representative list, for instance in analyses of protein architecture or the development of prediction methods.

How to get the list

The representative "25%-list," lists with decreasing levels of stringency (30%, 35%, 40% ... up to 90%), and the chains homologous to chains in the "25%-list" were stored in 1 data file ("pdb_select.dec_1993") and can be read from the Diskette Appendix or can be downloaded together with related data (DSSP secondary structures [Kabsch & Sander, 1983]; HSSP sequence families [Sander & Schneider, 1991]; and FSSP structure families [Holm et al., 1992]) using internet file transfer from ftp.embl-heidelberg.de (192.54.41.33). Log in as "anonymous," give your e-mail address as password, and change to the directory /pub/databases/protein_extras/pdb_select. Alternatively, the list can be requested from the EMBL file server netserv@embl-heidelberg.de (send the message text: "help proteindata"). In the future, the list may become available directly from the PDB.

The representative list will be updated on a regular basis (about twice a year). The coordinate entries corresponding to the list can be downloaded via internet directly from the PDB at pdb.pdb.bnl.gov (Brookhaven National Laboratories).

Note added in proof

Note of caution: Models built by homology or theoretical methods are not always clearly assigned in the current PDB. We have attempted to identify hypothetical models using the "EXPDTA" and comment fields and to exclude them from the list.

Acknowledgments

We thank our colleagues in the Protein Design Group for software tools and discussions and the Computer and Data Library Groups for collaborative support. Financial support from the Human Frontiers Science Program and the EC Bridge program was essential.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci 1*:409–417.

Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. 1992. A database of protein structure families with common folding motifs. *Protein Sci 1*:1691–1698.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers 22*:2577–2637.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Sander C, Schneider R. 1991. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet 9*:56–68.

Smith TF, Waterman MS. 1981. Comparison of biosequences. *Adv Appl Math 2*:482–489.