CABIOS

Vol. 12 no. 6 1996
Pages 525–529

The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value

R.W.W.Hooft, C.Sander, M.Scharf and G.Vriend

Abstract

Motivation: *The Protein Data Bank currently contains more than 4700 protein coordinate sets. It is often desirable to make a selection from these files based on a criterion like R-factor, experimental method, length of the amino acid sequence, or the number of homologous sequences in* SWISSPROT*. Doing this using the distributed form of the Protein Data Bank can be a tedious task, because (1) this requires reading one file for every single entry, and (2) not all of the information is present in a consistent computer readable way in all of the entries.*

Results: *The* PDBFINDER *database provides an easy to interpret file containing summary information about all Protein Data Bank files. Summary information from the* DSSP *(Definition of Secondary Structure of Proteins) and* HSSP *(Homology derived Secondary Structure of Proteins) databases is also included. Furthermore, where essential data were missing from the Protein Data Bank file, this information has been retrieved from the original literature.*

Availability: *The latest version of the* PDBFINDER *database can be downloaded by anonymous ftp from* `swift.embl-heidelberg.de`*, directory:* `/pdbfinder`.

Contact: *E-mail address* `hooft@embl-heidelberg.de`

Introduction

The Protein Data Bank (PDB) (Bernstein *et al.*, 1977) has become an invaluable resource of three-dimensional protein structures. Currently more than 4700 data sets are available, and the number is expected to double every 18 months. A few years ago, whenever a subset of available structures was needed that satisfied some simple criterion, one could visit a local expert that knew all published structures. With the current size of the database, these experts are no longer available everywhere. To satisfy the current requests for selections of data sets from the PDB the PDBFINDER database was created. This database contains summary information for all PDB data sets, extracted from the PDB, DSSP (Definition of Secondary Structure of Proteins, Kabsch and Sander, 1983), and HSSP (Homology derived Secondary Structure of Proteins, Sander and Schneider, 1991) databases. The extraction process is carried out automatically by the database update procedure at EMBL whenever any of the constituting databases is changed. This ensures that the information is always up to data.

During the extraction process, extensive sanity checking is performed to improve the accuracy of the data. Literature studies are carried out to fill gaps in the information present especially for older PDB files.

All information in the PDBFINDER database is generated by a PERL (Practical Extract and Report Language) (Wall and Schwartz, 1992) program, and represented in an ASCII file.

Many of the extractions are far from easy, due to more than 20 years of evolution of the PDB file format. All PDB files, including older ones containing much of the information in a less structured way, need to be recognized. A few examples of the extraction problems that we encountered are:

- For the 'Enzyme-Code' field the program contains seven different regular expressions (text patterns) that have to be tested sequentially. This is needed to recognize expressions like

  ```
  (E.C. 3.2.1.27)
  ```

  and

  ```
  EC: 3.2.1.27;
  ```

  as the same reference.
- For the extraction of the names of the programs used for refinement a set of regular expressions did not suffice. Here, in addition to the pattern recognition, the extracted information is compared with a list of well-known refinement programs. Finally, if no programs are found the 'REMARK 3' text is scanned for authors of the most well known programs. For example: '*The program by Konnert and Hendrickson*' is translated to NUCLSQ or PROLSQ depending on the context.
- In the 'HET' groups containing information on bound molecules, the name of the compound is often given as 'SEE REMARK [some number]'. This value is recognized, and the remark is scanned for the compound name instead using nine different regular expressions and a list of exceptions. This way the molecule name can be extracted from text like:

  HET GROUP SYSTEMATIC NAME: 5-(2-IMIDAZOLINYL)-2[2-(4-

Table I. Fields in the PDBFINDER database

ID The PDB-identifier, at the same time the unique identifier for each entry in the PDBFINDER database. A version of the PDBFINDER database is coupled to a PDB release: every PDB entry has exactly one corresponding record in the PDBFINDER.

Header The text from the HEADER record of the PDB file.

Date Subfield of *Header*. The date from the HEADER record in the PDB file, converted to a yyyy-mm-dd format (such that alphabetical sorting results in chronological order).

Compound Information from the COMPND records in the PDB file. Converted to mixed upper/lower case using the rules given in the PDB file specification.

Enzyme-Code Enzyme-code extracted from the COMPND records.

Source Information from the source records from the PDB file, converted to mixed upper/lower case.

Expr-Sys Expression system if given in the SOURCE records.

Author The authors of the PDB file, extracted from the AUTHOR records. Each author is mentioned in a separate 'Author' field. Converted to mixed case like 'Compound'.

Exp-Method The experiment used. This contains either 'NMR', 'X', 'FIBER', 'NEUTRON', 'MODEL', or 'OTHER'. This information is extracted from the EXPDTA field if present in the PDB file, otherwise it is dervied from the REMARK text.

Resolution Subfield of Exp-Method, only present for X-ray structures. Contains the Resolution of the data as given in 'REMARK 3'.

R-Factor Subfield of Exp-Method, only present for X-ray structures. Constains the R-factor of the data as given in 'REMARK 3', or as retrieved from the original literature.

Free-R Subfield of R-Factor, only present for a number of recent X-ray structures. Retrieved from 'REMARK 3'.

N-Models Number of 'MODEL's given in the PDB file.

SF-Type Subfield of R-Factor. The type of structure factor file, if one is available. This is either 'CIF', 'PDB' or 'UNKNOWN'.

Ref-Prog The names of all programs used in the refinement, separated by slashes. Only known programs (Table 2) will be listed. The information is extracted from 'REMARK 3' in the PDB file.

HSSP-N-Align Number of different SwissProt sequences in the HSSP file.

T-Frac-Helix Fraction of all protein residues in the structure that have 'helical' secondary structure.

T-Frac-Beta Fraction of all protein residues in the structure tha have 'strand' secondary structure

T-Nres-Prot Total number of protein residues in the structure.

T-non-Std Total number of non-standard protein residues in the structure.

T-Nres-Nucl Total number of nucleic acid residues in the structure.

T-Water-Mols Total number of water molecules given in the structure.

HET-Groups Number of HET-groups (normally drug molecules or metal atoms) present in the structure.

Het-Id Subfield of HET-groups, repeated for each HET-group. The unique identifier of the group.

Natom Subfield of Het-Id. Number of atoms in the HET-group.

Name Subfield of Het-Id. Name of the HET-group.

Chain Repeated for each protein or nucleic acid chain. The one-character chain identifier. For a chain without chain identifier, a '_' character is given.

Sec-Struc Subfield of Chain. Number of residues in this chain that have a secondary structure assignment. This field, and all its subfields, are extracted from the DSSP database. See Kabsch and Sander (1983) for more complete descriptions of this field and the subfields below

Helix Subfield of Sec-Struc. Number of residues in this chain that have a 'helix' secondary structure according to DSSP.

i,i + 3 Subfield of Helix. Number of residues in this chain in tight 3/10 Helices.

i,i + 5 Subfield of Helix. Number of residues in this chain in loose 5/18 Helices.

Beta Subfield of Sec-Struc. Number of residues in this chain that have a 'strand' secondary structure according to DSSP.

B-Bridge Subfield of Beta. Number of beta bridges observed

E-Beta Subfield of Beta. Number of residues in the chain that are in extended beta conformation

Para-Hb Subfield of Beta. The number of parallel strand backbone hydrogen bonds observed in the chain.

Anti-Hb Subfield of Beta. The number of anti-parallel strand backbond hydrogen bonds observed in the chain.

Amino-Acids Subfield of Chain. The number of protein residues in this chain.

non-Std Subfield of Amino-Acids. The number of non standard amino acids in the chain.

Miss-BB Subfield of Amino-Acids. The number of residues that have an incomplete set of backbone coordinates.

Miss-SC Subfield of Amino-Acids. The number of residues that have an incomplete set of side chain coordinates.

Table I. Fields in the PDBFINDER database. Continued

only-Ca Subfield of Amino-Acids. The number of residues that have only the C-alpha atom coordinates given.

UNK Subfield of Amino-Acids. The number of residues that is given as being of unknown type.

CYSS Subfield of Amino-Acids. The number of cysteine residues involved in S-S bridges.

Break Subfield of Amino-Acids. Number of chain breaks in the amino-acid chain.

Nucl-Acids Subfield of Chain. Number of Nucleic acid residues in the chain.

Substrate Subfield of Chain. Total number of atoms in substrates that are associated with this chain.

Water-Mols Subfield of Chain. Total number of water molecules that are associated with this chain.

Sequence Subfield of Chain. The sequence of the amino acid or nucleic acid residues in the chain, given as one-letter codes. Nucleic acids are given in lower case, amino acids in upper case.

HYDROXYPHENYL) 5-BENZIMIDAZOLYL] BENZIMIDAZOLE

This parsing can not solve all problems: some remarks assume more background knowledge, e.g.:

CHROMOMYCIN BINDS AS A MG2 CATION-COORDINATED DIMER EACH CHROMOMYCIN IS REPRESENTED AS A SET OF SIX HET GROUPS: BRI - ARI - CPH - CDR - CDR - ERI. THE COORDINATING MG IS PRESENTED AS A SEPARATE HET GROUP.

A number other fields are extracted using similarly elaborate schemes, and verified against expectable values. However, for those fields in the PDB file that have historically been created as plain english text by humans (*e.g.* 'SOURCE', 'COMPOUND', and 'REMARK 3') the automatic interpretation and correction efforts can be extended virtually without bounds. In such cases our work has concentrated on those pieces of information that can aid in navigation between databases and/or are often used as selection criteria for biocomputing purposes.

Many consistency checks can be performed based on the PDBFINDER information. Often, however, it is not possible to correct detected problems. Two example checks performed by us are author names and enzyme codes:

- A list of all author names was made where the same last name occurs with different initials, or two last names differ in one position only. Most of these pairs represent the same author. This list has been sent to the PDB data center for verification.
- All references from PDB to ENZYME (Bairoch, 1996) in the PDBFINDER were cross-verified with the references from ENZYME to PDB in the SWISSPROT (Bairoch and Apweiler, 1996) sequence database. All discrepancies have been sent to the PDB and SWISSPROT data centers for verification.

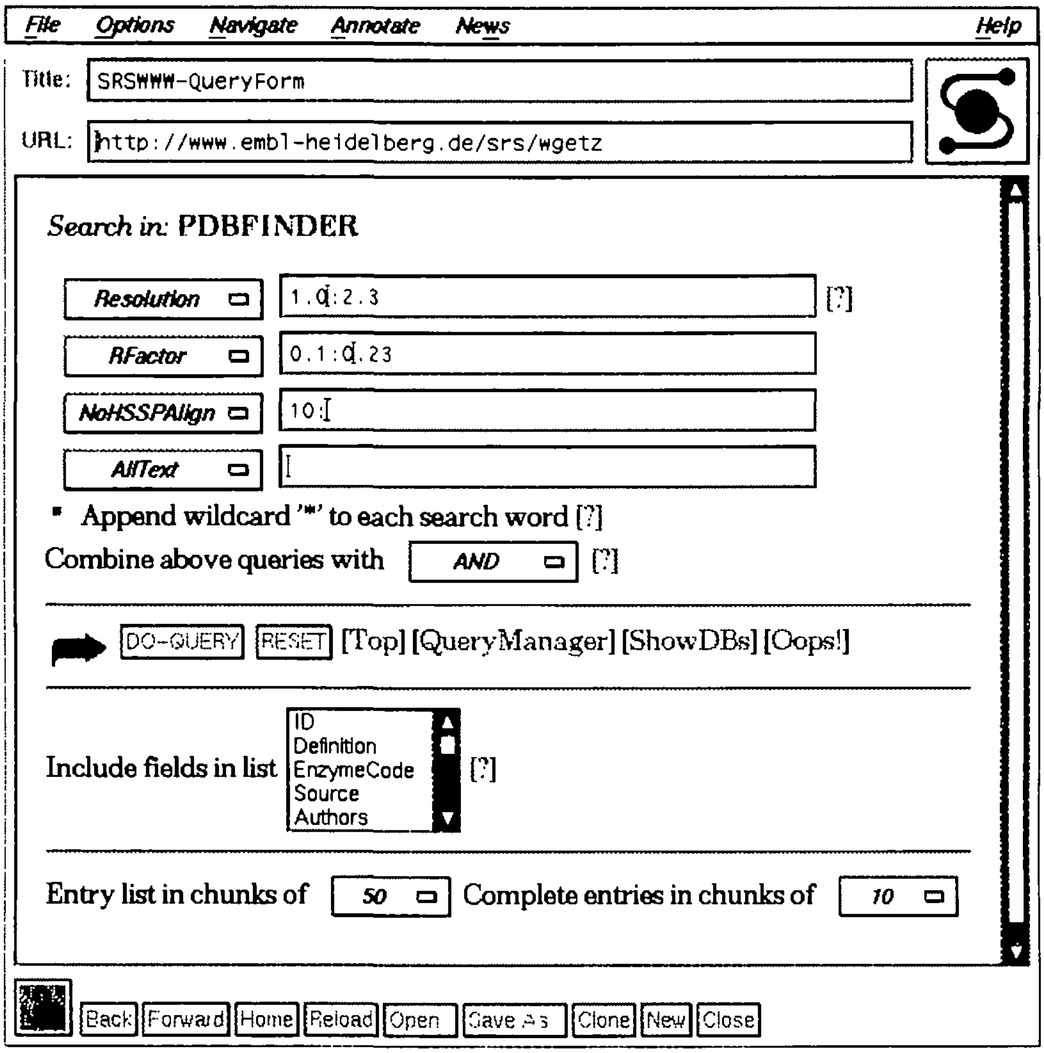
Table II. Recognized refinement programs

'AMBER', 'AMORE', 'ARP', 'ATOM', 'CALIBA', 'CEDAR', 'CHARMM', 'CORELS', 'CORMA', 'CRLS', 'CRYLSQ', 'DERIV', 'DGII', 'DIAMOND', 'DIANA', 'DINOSAUR', 'DISCOVER', 'DISGEO', 'DISMAN', 'DSPACE', 'ECEPP', 'EMBOSS', 'EREF', 'FANTOM', 'FREF', 'FRODO', 'GENERATE', 'GPRLSA', 'GRINCH', 'GROMOS', 'GROMOS-MDX', 'HABAS', 'HAFFIX', 'HKSCAT', 'INSIGHTII', 'IRMA', 'JACK-LEVITT', 'LOOP', 'MANOSK', 'MARDIGRAS', 'MIDGE', 'MM', 'MODELFIT', 'MUMOD', 'NCS', 'NOEMOL', 'NUCLIN', 'NUCLSQ', 'OMIT', 'OPAL', 'PIKSOL', 'PRESTO', 'PROFFT', 'PROLSQ', 'PROTEIN', 'PROTIN', 'PSFRODO', 'QUANTA', 'RESLSQ', 'RESTRAIN', 'ROTLSQ', 'RSREF', 'SCATT', 'SFALL', 'SFRK', 'SHELX', 'SHELXL', 'STEREOSEARCH', 'TNT', 'TOM', 'ULTIMA', 'VEMBED', 'X-PLOR', 'XEASY', 'YASAP'.

Since the people responsible for the databases have been

Table III. Example PDBFINDER entry for PDB file 1CBN (Teeter *et al.*, 1993). An illustration of the file format

//	
ID	: 1CBN
Header	: PLANT SEED PROTEIN
Date	: 1991-10-11
Compound	: Crambin
Source	: Abyssinian Cabbage (Crambe Abyssinica) Seed
Author	: M.M.Teeter
Author	: S.M.Roe
Author	: N.H.Heo
Exp-Method	: X
Resolution	: 0.83
R-Factor	: 0.11
Ref-Prog	: PROLSQ
HSSP-N-Align	: 9
T-Frac-Helix	: 0.35
T-Frac-Beta	: 0.11
T-Nres-Prot	: 46
HET-Groups	: 1
Het-Id	: 66
Natom	: 5
Name	: ETHANOL
Chain	: _
Sec-Struc	: 46
Helix	: 16
Beta	: 5
B-Bridge	: 1
Anti-Hb	: 6
Amino-Acids	: 46
CYSS	: 6
Substrate	: 5
Sequence	: TTCCPSIVARSNFNVCRLPGTPEALCA
	: TYTGCIIIPGATCPGDYAN



Title: SRSWWW-QueryForm

URL: http://www.embl-heidelberg.de/srs/wgetz

Search in: PDBFINDER

Resolution 1.0:2.3 [?]

RFactor 0.1:0.23

NoHSSPAlign 10:

AllText

- Append wildcard '*' to each search word [?]

Combine above queries with *AND* [?]

Include fields in list: ID, Definition, EnzymeCode, Source, Authors [?]

Entry list in chunks of *50* Complete entries in chunks of *10*

Fig. 1. An example PDBFINDER query using SRS: requested is a list of all PDB entries with an R-factor between 10% and 23%, a resolution between 1 0Å and 2.3Å, and 10 or more homologous protein sequences in the SWISSPROT sequence database.

informed, future versions can be expected to show less inconsistencies.

The PDBFINDER file format has been designed for high flexibility, and is readable from many programming languages (Search programs in awk, perl, C and Fortran have been used). Pseudo-hierarchical ordering of the data makes representation using an object-oriented paradigm easy.

A description of the fields in the PDBFINDER database is given in Table I, a list of currently recognized refinement programs in Table II, and an example entry in Table III.

Availability

FTP

The latest version of the PDBFINDER database can be retrieved by anonymous ftp from the EMBL server `ftp.embl-heidelberg.de`, in the directory

`/pub/databases/protein_extras/pdfinder`

or in compressed form from `swift.embl-heidelberg.de`, in the directory

`/pdbfinder`

This file is automatically regenerated whenever any of the HSSP, DSSP or PDB databases, which are also available from the EMBL server, are updated.

WHAT IF interface

The PDBFINDER database can also be accessed through the WHAT IF (Vriend, 1990) program; the SELECT menu in WHAT IF provides a PDBFINDER interface between the WHAT IF relational database of sequence-unique proteins and the Protein Data Base.

World Wide Web

The most convenient user interface to querying the PDBFINDER (and many other databases in molecular biology) is provided by Thure Etzold's SRS (Etzold and Argos, 1993) system, available via the World Wide Web on the internet via URL:

```
http://www.embl-heidelberg.de/srs/srsc
```

The PDBFINDER is currently indexed on the SRS server in Heidelberg and on eight other servers. In Heidelberg it can be found as one of the 'other libraries'. An example PDBFINDER query using SRS is shown in Figure 1.

The described work was done partly in the context of the 'protein structure verification project' funded by the European Commission. This project has as one of its goals to automate the process of data entry and validation for protein structure databases.

Acknowledgements

We would further like to acknowledge Karine Kromoian and Ingrid Warny for literature research, Thure Etzold for stimulating discussions, Hans Döbbeling and the EMBL computer group for continuous support, Antoine de Daruvar for database maintenance, and Enrique Abola and Amos Bairoch for their quick feedback on our questions.

References

Bairoch,A. (1996) The enzyme data bank in 1995. *Nucleic Acids Res.*, **24**, 221–222.

Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.*, **24**, 21–25.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer Jr,E.F., 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.

Etzold,T. and Argos,P. (1993) SRS - an indexing and retrieval tool for flat file data libraries. *Comput. Applic. Bio. Sci*, **9**, 59–64.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure· pattern recognition of hydrogen bond and geometrical features. *Biopolymers*, **22**, 2577–2637.

Sander,C. and Schneider,R. (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Teeter,M.M., Roe,S.M., and Heo,N.H. (1993) Atomic resolution (0.83 angstroms) crystal structure of the hydrophobic protein crambin at 130 K. *J. Mol. Biol.*, **230**, 292–311.

Vriend,G. (1990) WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.*, **8**, 52–56.

Wall,L. and Schwartz,R.L. (1992) *Programming perl.* O'Reilly & Associates, Sebastopol, CA, USA.

Received on May 21, 1996, revised and accepted on September 2, 1996