Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

PDB-Metrics: a Web tool for exploring the PDB contents

Renato Fileto, Paula R. Kuser, Michel E.B. Yamagishi, André A. Ribeiro, Thiago G. Quinalia, Eduardo H. Franco, Adauto L. Mancini, Roberto H. Higa, Stanley R.M. Oliveira, Edgard H. Santos, Fabio D. Vieira, Ivan Mazoni, Sergio A.B. Cruz and Goran Neshich

Embrapa Information Technology, Campinas, SP, Brasil
Corresponding author: G. Neshich
E-mail: neshich@cbi.cnptia.embrapa.br

Genet. Mol. Res. 5 (2): 333-341 (2006)
Received November 4, 2005
Accepted March 14, 2006
Published June 8, 2006

ABSTRACT. PDB-Metrics (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/index.html) is a component of the Diamond STING suite of programs for the analysis of protein sequence, structure and function. It summarizes the characteristics of the collection of protein structure descriptions deposited in the Protein Data Bank (PDB) and provides a Web interface to search and browse the PDB, using a variety of alternative criteria. PDB-Metrics is a powerful tool for bioinformaticians to examine the data span in the PDB from several perspectives. Although other Web sites offer some similar resources to explore the PDB contents, PDB-Metrics is among those with the most complete set of such facilities, integrated into a single Web site. This program has been developed using SQLite, a C library that provides all the query facilities of a database management system.

Key words: Protein Data Bank (PDB), PDB data distribution statistics, Search and recovery of the PDB contents

INTRODUCTION

The Protein Data Bank - PDB (Berman et al., 2000) is a collection of protein structures publicly available to the research community. Various laboratories around the world use the PDB and have produced PDB-derived databases and software packages, including the Diamond STING suite of programs for the analysis of protein sequence, structure and function (Neshich et al., 2004, 2005). Due to this widespread use and importance of the PDB, there is a growing demand for tools to help explore its contents.

The deciphered protein structures deposited in the PDB are described by files in the PDB format. The contents of such a file include, among other information, the coordinates of the atoms in the protein described by the file. A PDB file can be characterized by its size, its deposition date and a variety of characteristics of the corresponding protein, such as the number of chains, the number of models (for proteins resolved using nuclear magnetic resonance, NMR), the resolution (for non-NMR files), the number of atoms, and the frequency of each kind of residue in the protein structure.

PDB-Metrics is a database of measures extracted from the PDB collection of protein descriptions (the PDB files) and associated metadata (data used to describe data), with a Web interface for characterizing, browsing, searching, and recovering the PDB information. It enables sophisticated analysis and search of the PDB contents, by various perspectives, including listings of PDB files deposited each year, and grouping these files by their size and distinct characteristics of the protein structures. The reports generated by PDB-Metrics provide statistics about the distribution of diverse measures in the PDB collection. PDB-Metrics also provides an advanced search mechanism for posing queries involving the measures maintained in its database.

There are several tools available on the Web for exploring the contents of the PDB, including the PDB Web site itself (Deshpande et al., 2005). Nevertheless, from the best of our knowledge, PDB-Metrics has the richest repertoire of criteria for searching, browsing and analyzing the PDB contents on the Web nowadays. A user can benefit from the PDB-Metrics while developing custom software applications in areas such as molecular graphics, structure analysis and verification and simulation. The remainder of this paper describes the PDB-Metrics Web interface, its facilities for exploring the PDB contents, and some implementation issues.

THE PDB-METRICS WEB INTERFACE

The main page of PDB-Metrics, which can be accessed through the URL http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/index.html, shows at the top the number of PDB, HSSP and Prosite entries catalogued in the PDB-Metrics database. This database is updated on a weekly basis. The date and time of the most recent update of the PDB-Metrics database appear on the main page, just below the indication of the number of entries. The date and time refer to the moment at which PDB-Metrics started the updating process, using a local copy of the PDB present in our server, which has usually been synchronized with the PDB central repository some minutes earlier.

The search and browsing facilities provided by PDB-Metrics are described below. Most of the reports generated by these facilities include statistics of the data occurrences in the PDB collection.

Search by name

Recovers a specific PDB structure by the protein name, allowing access to all the data files available in our server for that particular protein (e.g., **.PDB, *.HSSP, *.CON (Contacts File), *.AIR (Accessibility and Interface Residues), *.ANG (Dihedral Angles File)*) (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/PDB_byName.htm). In the report generated by PDB-Metrics, one can follow a link to see the description of the active site(s) of the structure, provided by the European Bioinformatics Institute (http://www.ebi.ac.uk/), and visualize the structure of the protein using STING.

PDB files obtained by NMR

Lists the PDB structures obtained by NMR. The loading of the complete report of NMR files can take a while, because of its size. For faster access, one can click on the link *Show the PDB list by name*, in order to see only the list of PDB names (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/NMR.html).

PDB files by type

Lists the types of PDB structures taken from the *Header* section of all deposited PDB structures. The types are presented in alphabetic order. For each type, the program presents the number of structures of that type in the PDB collection and allows the user to check the protein structures of that type by pressing the respective link (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/family.html).

PDB files by deposition year

Shows the number of PDB structures deposited each year and allows the user to see the list of files deposited in a particular year by following a link (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/deposition.html).

PDB files by size

Lists the deposited PDB structures grouped in ranges of file size (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/size.html).

PDB files by number of models

Lists the PDB structures grouped by their number of models. This option takes into account only the structures obtained by NMR (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/nr_models.html).

PDB files by number of chains

Lists the PDB structures grouped by their number of chains (http://sms.cbi.cnptia.

embrapa.br/SMS/pdb_metrics/nr_chains.html).

PDB files by number of DNA chains

Lists the PDB structures grouped by their number of DNA chains (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/nr_DNA_chains.html).

PDB files containing chains with no identifier

Lists the deposited PDB structures containing at least one chain with no identifier (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/no_identifier.html).

PDB files by resolution

Lists the PDB structures grouped according to their resolution. The resolution is not applicable for structures obtained by NMR (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/resolution.html).

PDB files by number of identified H_2O molecules

Lists the PDB structures grouped by their number of crystallized water molecules (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/crystalHOH.html).

PDB files by number of identified ligands

Lists the PDB structures grouped by the number of identified ligands (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/ligands.html).

PDB files by number of identified residues

Lists the PDB structures grouped by the number of identified residues (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/nr_id_residues.html).

PDB files by number of atoms

Lists the PDB structures grouped by their total number of atoms (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/nr_atoms.html).

Curiosities (extreme and average values in PDB)

Presents the minimum, the average and the maximum values of the following measures across the whole PDB database: file size, number of chains, number of DNA chains, number of water molecules, number of identified ligands, number of identified residues, number of atoms, resolution, and number of models (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/curiosities.html).

Residue frequency for all PDB files

Presents a histogram of the frequency of each residue or residue family across all structures deposited in the PDB (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/frequency.html).

Advanced search

Allows querying the PDB-Metrics database using diverse criteria based on the values of the measures described above and keywords occurring in the *Header*, *Title*, *Compound*, *Source* and *Keywords* sections of the PDB files (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/AdvSearch.htm).

Figure 1 presents the PDB-Metrics main menu (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/index.html) (on the left) and the results generated by two of its options: Curiosities (extreme and average values in PDB) (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/curiosities.html) and Residue Frequency for all PDB files (http://sms.cbi.cnptia.embrapa.br/SMS/pdb_metrics/frequency.html), just to illustrate the PDB-Metrics functionality.

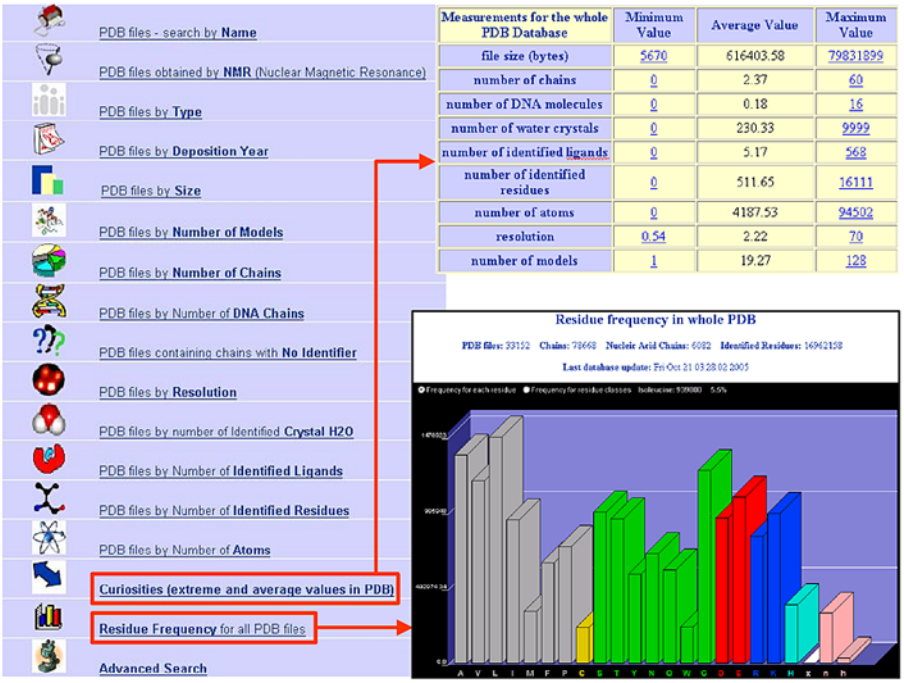

Figure 1. The main menu of PDB-Metrics (left). The extreme values (top right), with the non-weighted average values in the middle column. The residue frequency histogram for the whole PDB collection (bottom right), showing that leucine (L) is the most prevalent residue, with 1,433,536 occurrences, making 8.7% of all residue occurrences in the whole PDB collection, while the partially identified residues (x) are just 76, representing a very small portion of the 16,962,158 identified residues. These results were obtained on October 24, 2005.

THE ADVANCED SEARCH FACILITY

The PDB-Metrics' advanced search facility allows the user to specify a search on the PDB contents by filling out the search criteria in a form. Figure 2 presents the specification of a query intended to recover PDB files containing the structures of complexes of proteins, peptides, viruses, or DNA, deposited since the year 2000, whose molecular structure was determined by X-rays, with resolution between 0.5 and 1.8 angstroms and having at least 2 chains, 2 crystallized water molecules and 1 ligand per file. Only the structures satisfying all the provided criteria will be recovered. A user is encouraged to supply enough criteria to restrict a query the most, in order to get a fast and short response.

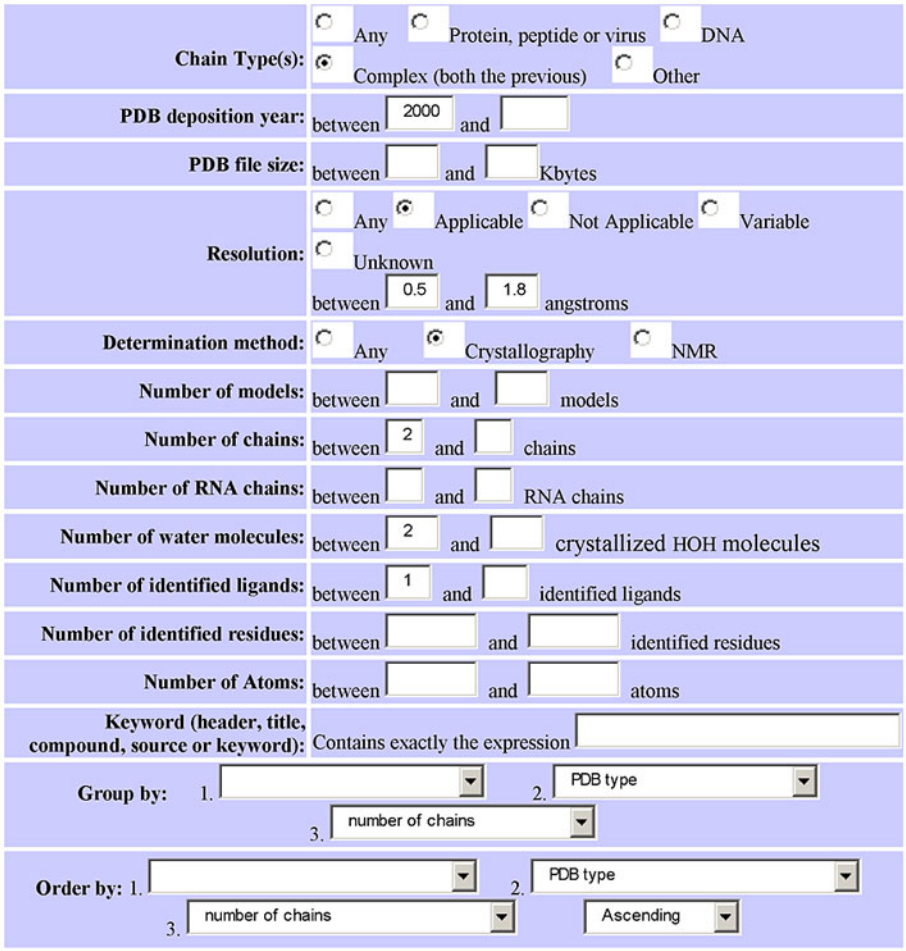

Figure 2. Posing a query on the PDB-Metrics' advanced search specification form.

The *Group by* and the *Order by* sections of the advanced search specification form enable the specifications of the attributes used to group and sort the results. The values supplied in these two sections must be consistent. That is, if some criteria are chosen to group the results,

the same criteria are used to order those results, because of limitations of the query processing engine.

The criteria inserted in the form above are translated into an SQL query (Date, 1993) which is then submitted to the PDB-Metrics database engine. Figure 3 shows the SQL query corresponding to the query specified in the form presented in Figure 2.

```
SELECT   pdb_type , nr_chains , count(*)
FROM   pdb_file_mt
WHERE  PDB_chain_type = 'N'  AND  PDB_dep_year >= '2000'  AND
          resolution_type = 'A'  AND  determ_method = 'C'  AND
          nr_chains >= '2'  AND  nr_HOH >= '2'  AND  nr_id_ligands >= '1'
GROUP BY   pdb_type,   nr_chains
ORDER BY   pdb_type,   nr_chains;
```

Figure 3. An SQL query translated from the PDB-Metrics' advanced search specification form.

The result of the query applied to the database is presented as a list of items, each one defining a group of the PDB files satisfying the search. Figure 4 presents the results for the query specified in Figures 2 and 3 (clearly, the results may vary as the PDB collection is updated). In order to access the listings of the PDB files in a specific group (described by a line of the table in Figure 4), one can follow the link on the number of PDB files in that group (last column of the table in Figure 4).

pdb_type	nr_chains	# of PDB files
DNA BINDING PROTEIN/DNA	8	1
HYDROLASE/DNA	3	8
HYDROLASE/DNA	4	4
HYDROLASE/DNA	6	3
HYDROLASE/DNA	8	1
HYDROLASE/DNA	12	1
LYASE/DNA	12	1
PROTEIN/RNA COMPLEX	6	1
SIGNALING PROTEIN/RNA	2	3
TRANSCRIPTION FACTOR	4	1
TRANSCRIPTION/DNA	3	1
TRANSCRIPTION/DNA	4	2
TRANSCRIPTION/DNA	8	1
TRANSCRIPTION/RNA	4	2
TRANSFERASE/DNA	3	11
TRANSFERASE/DNA	6	2
TRANSPORT PROTEIN/DNA	2	2

Figure 4. The result of the query described in Figure 2 and Figure 3, generated on October 24, 2005: there are 45 PDB files satisfying the query, divided into 17 groups (each one defined by a value of the PDB type and a certain number of chains), which are ordered by the PDB type.

IMPLEMENTATION ISSUES

PDB-Metrics has been implemented using the SQLite database management tool (Hipp, 2005) to maintain its database and process queries. The major reason for this choice is that SQLite offers most of the functionality of a full-fledged relational database management system (Elmasri and Navathe, 2003), while being just a C library that requires virtually no configuration and maintenance. Furthermore, SQLite is a very efficient and highly portable database engine. These traits make SQLite the best current option for the implementation of PDB-Metrics, as it provides powerful data management and query facilities, while imposing minimal administration burden to run PDB-Metrics at the STING mirror sites running on heterogeneous platforms.

The Web interface of PDB-Metrics was written as CGI using the Perl programming language (Schwartz et al., 2005). The DBI application program interface (Descartes and Bunce, 2000) allows the use of SQL (Date, 1993) as an embedded language to access the SQLite database from the Perl programs.

Only advanced queries must be processed against the database for each user request. The other pages in the two top levels of the PDB-Metrics (the main page and the reports accessible from it) are built every time the database is updated and kept ready for future requests. These pages are generated and stored in the STING main site and replicated at the STING mirror sites, in order to optimize access and save computational power.

CONCLUSIONS

PDB-Metrics is a database that summarizes several characteristics of the PDB collection of protein structure descriptions. The PDB-Metrics Web interface offers a vast variety of facilities for analyzing, browsing, searching, and visualizing the PDB contents, according to various measures present in PDB files. The reports and graphics generated by PDB-Metrics group the PDB entries according to the values of different characteristics and measures of the protein descriptions, such as the file size, the type of protein, its number of chains and its number of residues of different types. These outputs include statistics of the span of these values in the whole PDB collection of protein structure descriptions. Although many of the PDB-Metrics facilities and features are provided by other programs, including the PDB site itself (Deshpande et al., 2005), PDB-Metrics offers all these resources in a single-integrated Web tool, being the most complete and functional tool for exploring the PDB contents currently available on the Web.

REFERENCES

Berman HM, Bhat TN, Bourne PE, Feng Z, et al. (2000). The Protein Data Bank and the challenge of structural genomics. *Nucleic Acids Res.* 28: 235-242.

Date CJ (1993). A guide to the SQL standard. Addinson-Wesley Publishing Company, Way Reading, MA, USA.

Descartes A and Bunce T (2000). Programming the Perl DBI: Database programming with Perl. O'Reilly & Associates Inc., Sebastopol, CA, USA.

Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, et al. (2005). The RCSB Protein Data Bank: A redesigned query system and relational database on the mmCIF schema. *Nucleic Acids Res*. 33: D233-D237.

Elmasri R and Navathe SB (2003). Fundamentals of Database Systems. 4th edn. Addison-Wesley, Menlo Park, CA, USA.

Hipp R (2005). SQLite: a small C library that implements a self-contained, embeddable, zero-configuration SQL database engine. Available at http://www.sqlite.org/ (Accessed September 2005).

Neshich G, Higa RH, Yamagishi MEB, Mancini A, et al. (2004). SMS: Integrated software for extensive analyses of 3d structures of proteins and their complexes. *BMC Bioinformatics* 5: 107.

Neshich G, Borro LC, Higa RH, Kuser PR, et al. (2005). The Diamond STING server. *Nucleic Acids Res*. 33: W29-W35.

Schwartz R, Phoenix T and Foy BD (2005). Learning Perl. 4th edn. O'Reilly & Associates Inc., Sebastopol, CA, USA.