

Evaluation of structure quality tutorial, using RCSB PDB tools

EMBO Lecture Course

3D Structure Databases- Uses for Biological Problem Solving

Workshop Leader: Kyle Burkhardt, RCSB PDB, Rutgers University.

In order to validate structural data using the RCSB PDB tools, the coordinate (and structure factor) files should be available to upload into the server for validation. These files may either be generated based on data that a user has collected or downloaded from the PDB. The following excerpts of a unit in the Current Protocols in Bioinformatics (submitted September 1, 2004) describe the procedures used for downloading structural data from the PDB followed by a description of the steps involved in running the RCSB Validation Server.

=====

Downloading structural data from the PDB

Downloading structural data from the PDB allows the user to visualize, analyze and modify the file(s) locally for their own research. Thus users may wish to download either one or many structures from the PDB. Structures may be downloaded using the Web download, FTP, or CD-ROMs that are distributed by the RCSB. If only a few files need to be downloaded the Web download is probably most suitable, while FTP or CD-ROMs are more appropriate when downloading large numbers of files (several hundreds). A set of structures which are not related to each other in some way (i.e. a set of structures that cannot be obtained via a query), may be downloaded using a script from the FTP site. The steps involved in downloading files using each of these methods are described here.

Necessary Resources

Hardware

A computer with Internet access is necessary for the Web download. A computer with FTP capabilities is needed for the FTP download. A computer with the programming language Perl installed is needed for running FTP download scripts.

Software

A Web browser, Program for FTP access (either via a Web browser or command line FTP client)

Files

None required for standard Web or FTP download. For scripted FTP downloads, the scripts `getPdbUpdate.pl` or `getPdbStructures.pl` may be downloaded from <ftp://ftp.rcsb.org/pub/pdb/software>.

Web downloading structural data from the PDB

1. Select the file(s) to be downloaded using any one of the Query methods (either from the RCSB PDB home page, or using QuickSearch, SearchLite or SearchFields).

For example if the PDB ID of the entry is known, enter it in the box provided on the RCSB PDB home page. This launches the Structure Explorer page from where the coordinates, experimental data or sequences may be downloaded. Alternatively, a query can be designed to select the group of structures that are to be downloaded. If the query returns more structures than need to be downloaded, select the desired structures using the check box on the left hand side of each entry in the Query Result Brower page.

2. Download the selected file(s).

When downloading a single file from the Structure Explorer page, click on the “Download/Display File” option on the left hand side of the Structure Explorer page. This opens a new page from where the file type and compression mode can be selected for the download. Either the PDB format (Callaway et al., 1996) or mmCIF format (Bourne et al., 1997) coordinate files may be downloaded with or without compression. The user is strongly recommended to use the uniformly formatted mmCIF files for most purposes since the information content and annotation of these files have been standardized and completed as a part of a major data uniformity project (Westbrook et al., 2002). Compressed XML format files may also be downloaded from this page. The data content of the XML format files is equivalent to the data content of the mmCIF format files. Experimental data can be directly downloaded from the left hand side of the Structure Explorer page by clicking on the word “compressed” (written either under “Structure factors” or “NMR constraints”). In addition, the coordinates of the complete biological unit(s) are also available from this page in compressed form. The biological assembly may be the same as the set of submitted coordinates, or it may be smaller or larger. A “biological unit” download is available in either case.

Multiple files can be downloaded from the Query Result Browser page by choosing the “Download Structures or Sequences” option from the pull down menu at the top of the Query Result Brower page. If all files in the Query Result Browser page do not need to be downloaded,

coordinate files of the appropriate structures can be selected for downloading. A new page is opened from where PDB or mmCIF format coordinate files or only the sequences of the selected structures can be downloaded.

Downloading structural data via FTP

1. Access the ftp site at `ftp://ftp.rcsb.org`, or click on the “DOWNLOAD files” link in the top left corner of the RCSB home page (`http://www.pdb.org`).

A variety of tools can be used to access the FTP site. Most browsers will allow a user to simply type the FTP address (`ftp://ftp.rcsb.org`) into the address bar. Many computers (in particular, those running Unix, Linux, Mac OSX, or related operating systems) will also have readily available command line FTP clients. For the latter option, a user would type “`ftp ftp.rcsb.org`” on the command line, enter “anonymous” for the user name, and the user’s email address for a password.

The organization of the FTP site is available in the README file included at the top level directory at this site. Some help files containing information about how to download files and a listing of the contents of the directories are also available.

2. Download the structure(s) following the instructions listed in the README file.

A sample FTP session to download the PDB entry 1O3Q (Chen et al., 2001) from the FTP site is as follows:

```
%ftp ftp.rcsb.org
```

```
Connected to ftp.rcsb.org.
```

```
...
```

```
Name (ftp.rcsb.org:user): anonymous
```

```
331 Guest login ok, send your complete e-mail address as password.
```

```
Password: ...
```

```
...
```

```
ftp> cd pub/pdb/data/structures/divided/mmCIF/o3/
```

```
250 CWD command successful.
```

```
ftp> binary
```

```
200 Type set to I.
```

```
ftp> get 1o3q.cif.Z
```

```
200 PORT command successful.
```

150 Opening BINARY mode data connection for 1o3q.cif.Z (88991 bytes).

226 Transfer complete.

local: 1o3q.cif.Z remote: 1o3q.cif.Z

88991 bytes received in seconds (.... Kbytes/s)

ftp> quit

Downloading structural data from the FTP archive via automated scripts

1. Download the Perl scripts `getPdbUpdate.pl` and `getPdbStructures.pl` from <ftp://ftp.rcsb.org/pub/pdb/software>.

Additional documentation for these scripts is available at <ftp://ftp.rcsb.org/pub/pdb/software/getPdbUpdate.html> and <ftp://ftp.rcsb.org/pub/pdb/software/getPdbStructures.html>

2. Run the scripts on the command line as described in the online documentation, *e.g.*:

```
getPdbUpdate.pl latest
```

```
getPdbUpdate.pl 20040816
```

```
getPdbStructures.pl my_list_of_pdb_ids.txt
```

This form of access is useful for downloading a set of structures which are not necessarily related to each other, for example a set of structures that cannot easily be obtained through queries on the Web page. The Perl scripts provided by the PDB will run on most computers that have Perl installed (including Windows, Unix, Linux, Mac OSX). However, these scripts have additional requirements which are detailed in the online documentation.

Accessing Structural data from the CD ROM

1. Subscribe to the CD-ROM distributions via an electronic form or by email using instructions provided at http://www.rcsb.org/pdb/data_cd.html.

The entire contents of the PDB including coordinate files, structure factor files, NMR constraints and some resources may be accessed using the CD-ROMs.

2. Access the file(s) and copy them to your local computer for visualization and analysis.

This form of access is useful where Internet access is not available.

Assessing the quality of a macromolecular structure using the Validation Server

The Validation Server is a public, Web-based resource that can be used to examine structures downloaded from the PDB in order to select PDB entries for structure-function studies, molecular modeling, and drug-design. It can also be used by depositors to track the progress of refinement of their structure or during the preparation of structural data for deposition to the PDB. The Validation Server performs two functions – Precheck and Validate. The Precheck function checks the format of the files uploaded for validation, while the Validate function checks the geometry, chemistry, sequence of the structure, and computes various derived features and reports from PROCHECK (Laskowski *et al.*, 1993) and NUCHECK (Feng *et al.*, 1998), that can be used to assess its quality. If a structure factor file is uploaded, SFCHECK (Vaguine *et al.*, 1999) reports are also generated. Major steps involved in using the Web-based Validation Server are described here.

Necessary Resources

Hardware

A computer with Internet access

Software

A Web browser, any standard Postscript viewer (like Ghostscript, Ghostview, GSview), RasMol (Sayle & Milner-White, 1995) or any other molecular visualization software that allows visualization of atoms and residues of the structure.

Files

For validation, the coordinate file must either be in standard mmCIF or PDB format. The mmCIF format files used for validation should either be downloaded from the PDB (to ensure that the file is complete and standardized) or generated by the **pdb_extract** (Yang *et al.*, 2004) applications (to ensure that the file conforms to the standard mmCIF dictionary). Both mmCIF and PDB format files must include information about the unit cell, space group and chain IDs. A summary of corresponding categories and remarks that are necessary for running validation on mmCIF and PDB format files respectively are shown in Table 1. These categories and remarks were taken for an example PDB entry 1O3Q (Chen *et al.*, 2001). If an alternate setting is being used for space group symmetry, appropriate matrices to transform the orthogonal coordinates to fractional

coordinates should be included in the file. For a more complete report, the coordinate file should also contain sequence information about all polymers (protein or nucleic acid) present in the structure; this information has been updated in all mmCIF format files downloaded from the PDB. Files that are being prepared for deposition may not yet have the sequence information. In this case, validation will still run but the list of residues that are extra or missing in the coordinates (compared to the sequence provided) will not be generated. If a structure factor file is included for validation, it must be in mmCIF format and must be validated along with its corresponding coordinate file. It is recommended that this file is compressed before uploading.

Running the Web-based Validation Server

1. Access the Web-based Validation Server at RCSB (<http://deposit.pdb.org/validate/>).
2. From the pull-down menus provided, select the experimental method (X-ray or NMR) and press the BEGIN button. This opens a new Web page.

Note that the Web-based Validation Server is in fact part of the ADIT tool that is used for depositing structures .

3. Enter the path/location of the coordinate file and structure factor file (if applicable) in the space provided. The files can also be located using the Browse button.
4. Specify the coordinate file type (mmCIF or PDB) and structure factor file type (mmCIF or other).

In the ADIT tool that is used to deposit structures, mmCIF or other ASCII format structure factor files can be uploaded. However, validation reports for structure factor files (using the Web-based Validation Server) can only be created for files submitted in mmCIF format.

5. Select the operation to perform (Precheck or Validate) and then press the Begin button. *It is recommended that the user should start with the “Precheck” option to make sure that the format of the uploaded file is correct. If the file format has been previously checked, then the user can select the Validate option here.*

7. After Precheck has been successfully completed, press “Continue” to proceed to the next page. Select “Validate” from the operation menu, and then press “Begin” to create the validation report.

The validation may take a few minutes depending on the size of the structure.

8. Review the validation reports presented in the browser window (Figure 11).

For an X-ray structure downloaded from the PDB (in mmCIF format) with its accompanying structure factor file, all the validation reports are accessible from a single Web page. These include an Atlas summary, with links to view molecular graphics of the asymmetric unit and crystal packing. Links to view molecular graphics of the biological unit for the entry are displayed in cases where the biological unit is either larger or smaller than the asymmetric unit. A validation summary letter is generated that includes structural diagnostics, listing close contacts within the asymmetric unit and between symmetry related molecules; bond length and angle deviations from standard dictionary values; chirality errors; list of waters that are more than 3.5 Å from either the protein or nucleic acid molecules present in the structure; residues and atoms that are missing from the coordinates due to disorder but were present in the sample whose structure was determined; alignment of sequence and coordinates if any residues were missing. Any peptide bond distances that are too long compared to the standard values are also identified here. Additionally, results from the validation programs PROCHECK, NUCHECK and SFCHECK are also reported.

If the structure was validated in preparation for deposition to the PDB, in addition to the reports described above, sequence-coordinate mismatches, extra residues, and lists of extra atoms in residues or ligands in the macromolecules may also be generated. Use Rasmol or other visualization software to visually check these atoms and residue(s) for a better understanding of the nature and source of deviation. These errors should be corrected before the files are deposited to the PDB.

Reference:

- Bourne, P.E., Berman, H.M., Watenpaugh, K., Westbrook, J.D. & Fitzgerald, P.M.D. 1997. The macromolecular Crystallographic Information File (mmCIF). *Meth. Enzymol.* 277:571-590.
- Callaway, J., Cummings, M., Deroski, B., Esposito, P., Forman, A., Langdon, P., Libeson, M., McCarthy, J., Sikora, J., Xue, D., Abola, E., Bernstein, F., Manning, N., Shea, R., Stampf, D. & Sussman, J. (1996). Brookhaven National Laboratory.
- Chen, S., Vojtechovsky, J., Parkinson, G.N., Ebright, R.H. & Berman, H.M. 2001. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: DNA binding specificity based on energetics of DNA kinking. *J. Mol. Biol.* 314:63-74.
- Feng, Z., Westbrook, J. & Berman, H.M. (1998). Report NDB-407. Rutgers University, New Brunswick, NJ.

- Laskowski, R.A., McArthur, M.W., Moss, D.S. & Thornton, J.M. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26:283-291.
- Sayle, R. & Milner-White, E.J. 1995. RasMol: biomolecular graphics for all. *Trends Biochem. Sci.* 20:374.
- Vaguine, A.A., Richelle, J. & Wodak, S.J. 1999. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr.* D55:191-205.
- Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., Bourne, P.E. & Berman, H.M. 2002. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res.* 30:245-248.
- Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H.M. & Westbrook, J. in press. Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr.*

Internet resources

Web-link	Description
http://www.pdb.org/	The home page for RCSB PDB
ftp://ftp.rcsb.org	Main RCSB PDB FTP site
http://deposit.pdb.org/validate/	Validation Server at the RCSB PDB