Protein Data Bank & Structure Deposition at PDBj

Nobutoshi Ito

Graduate School of Biomedical Science, Tokyo Medical and Dental University
also
Protein Bank Japan

http://www.pdbj.org

CCP4 Workshop
2 Nov 2012
Today’s menu

- wwPDB and PDBj
- Quick tour of deposition
- Quantity and quality of data
- Related databases etc at PDBj
E-MSD is supported by grants from the Wellcome Trust, the EU (TEMBLOR, NMRQUAL and IIMS), CCP4, the BBSRC, the MRC and EMBL.

The BMRB is supported by NIH grant LM05799 from the National Library of Medicine.

PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

The RCSB PDB is supported by grants from the National Science Foundation, National Institute of General Medical Sciences, the Office of Science-Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes & Digestive & Kidney Diseases.

wwPDB and wwPDBAC members at Rutgers Univ. on 1 Oct. 2010

Kleywegt, G    Berman, HM    Markley, JL    Nakamura, H
International collaboration in wwPDB

1) Curation, data processing, and registration are made by all the members, collaborating with each other.
2) We have a single data archive, which is looked after by one “archive keeper (RCSB)”.
3) Data format and new descriptions are discussed among the members.
4) Members are encouraged to develop their own browsers, viewers, and other APIs and services.

219 M files were downloaded from three wwPDB members (PDBj, RCSB-PDB, PDBe) during a year.
Protein Data Bank Japan (PDBj)

http://www.pdbj.org/

At Institute for Protein Research, Osaka Univ. since 2001.
Supported by the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST).
Protein Data Bank Japan: PDBj

1. The *wwPDB* management as one of the members
2. **Curation and data processing** for the deposited data (PDB and BMRB)
3. Data remediation and development of the **format** for correct description
4. Addition of experimental information from **literatures**
5. Development of **query tools and derived databases** as the web service
PDBj curaa Quarter of the deposited data, mainly from Asian and Oceania regions.
Quick Tour of How to Deposit (X-ray) Structures with ADIT
Before deposition, please check;

- Is it the **final** structure?
- Coordinates **and structure factors** (raw images in the future?)
- You have a sequence file? (**real** sequence used in the experiment, one-letter code, database ID… If it’s a new protein, sequence deposition is also required)
- Ligand information (IUPAC name, images…)
To make the procedure easier

- Have various log files ready (crystallisation note, data collection, scaling...)
- Run preprocessing programs (pdb_extract, deposit_mmcif, CCP4 data harvesting...)
How to Deposit

PDB Deposition

NMR Data Deposition

What's new

6-Oct-2010
Effective December 6, 2010, deposition of chemical shift data will be mandatory when submitting NMR entries to the PDB. (more...)

19-Aug-2010
Validation Report PDFs (more...)

13-Jul-2010
The latest version of JV (v3.8), the molecular graphics viewer developed by PDBj, has been released. A new functional command "displayatom on/off" has been added, which provides various display methods easily. Please download the newest version from here.

30-Jun-2010
Version 2 NMR restrain files have been released. (more...)

19-May-2010
The latest version of JV (v3.7.1) has been released.

7-May-2010
The IP address of PDBj ftp server (ftp.pdbj.org) has been changed to 133.1.158.142. If you access to the PDBj ftp using the IP address, directly, please change the IP address with the URL (ftp.pdbj.org) to access PDBj Site.
PDBj Validation & Deposition Portal

Welcome to the PDBj validation/deposition portal.

Need help?

If you have any questions, please read “Deposition tutorial” first. If you still need help, please contact us. Questions, comments, and suggestions in validation/deposition service and tutorial are welcome.

Data Deposition Tools

Search

You may search HET groups using Ligand Expo. If your HET groups are NOT new, you should use the existent ones.

Check sequence database references for proteins or nucleic acids in your structure at UniProtKB. The author is encouraged to submit their sequence information via SPIN (the UniProtKB submission tool) for directly sequenced proteins, or vir submission tools provided by INSDC members (SAKURA at DDBJ, Webin at EMBL, BankIt at GenBank) if the nucleotide sequence is available.

Validate

Before you deposit the structure data, validate your structure using the validation server at PDBj.

Deposit

After you validate your structure data, deposit your structure using the structure deposition tool ADIT at PDBj, beta-ADIT or ADIT-NMR at PDBj (for NMR studies). For other deposition servers at wwPDB member sites, please visit from the wwPDB web site.
プロテインデータバンクへの構造解析データ登録のご案内

PDB (Protein Data Bank)への登録は、ウェブブラウザを利用して対話的に登録を行うことができるADIT (Auto Deposition Input Tool)を使用して行います。また、必要に応じてADITを使用すると、ADITで手動入力する項目を深くすることができます。このページでは登録の手順、必要なデータ類とその形式や注意事項などについてご案内いたします。

本チュートリアルについて（ご利用前に必ずご覧ください）
 登録に必要な流れ
 登録時の確認事項
 予備データファイルの準備
 様々なファイルの準備
 登録情報の抽出
 登録データの検証

PDBj

（C）PDB
ADIT Deposition Tool

Need help?

If you have any questions, please read “Deposition tutorial” first. If you still need help, please contact us. Questions, comments, and suggestions in validation/deposition service and tutorial are welcome.

If your question is about a particular entry, please include the RCSB ID in your message.

Announcements

- If TLS is involved in refinement, Please read “Depositors for using TLS” document.
- Please make sure that your depositing coordinate meets wwPDB policy. wwPDB annotation POLICIES document is available from wwPDB website in HTML or PDF format.
- Structure factor amplitudes/intensities (for crystal structures) and restraints (for NMR structures) are a mandatory requirement for PDB deposition. This policy is published at wwPDB news.
- Structures solved by Electron Microscopy should deposit their maps to the EMDB (at PDB or RCSB) first, and should supply the EMD-id as an associated entry with their ADIT deposition.

Files and information required to deposit X-ray, NMR, and Electron Microscopy, or other methods structures.

Start New Session

To start a new ADIT session, select the experimental method and the molecular structure type. Then press the BEGIN button.

| Method: X-ray | Structure Type: Protein | BEGIN |

Continue Previous ADIT Session

To continue a session from an earlier date, Enter your session restart ID and press the CONTINUE SESSION button.

<table>
<thead>
<tr>
<th>Session Restart ID:</th>
<th>CONTINUE SESSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>(example: 2006-07-01.serverhost1name.9999.12345678)</td>
<td></td>
</tr>
</tbody>
</table>

If you have started an ADIT-NMR session, you can continue at ADIT-NMR.
**Running an ADIT Session**

1. Please review the format requirements before starting the session.
2. You may wish to use pdb_extract to process log files for deposition.
3. Check ligands with Ligand Expo.

**About ADIT**

An ADIT session involves three steps:

1. A data format precheck
2. A validation check
3. The actual deposition of the structural data to the database

**Precheck** confirms that the data files you enter are in a format that can be automatically processed by our software. The format and required data items for coordinate files are described for the PDB and mmCIF formats. The required mmCIF data items for structure factor data are also described. The precheck identifies any changes that need to be made in your data files in order to obtain a validation report or to deposit your structure.

The **validation check** creates a validation report for the structure which includes an RCSB PDB Atlas page; a validation report with stereochemical checks, and, depending on the type and content of the files uploaded, produces Precheck, Molprobity, NUCheck and SFCheck reports. The validation may take a few minutes depending on the size of your...
Coordinate format is OK!

Sequence information:

One polymer chain was extracted from the coordinate section. The detailed residue information of each chain is listed below. Standard residues are listed as one letter code sequence. Non-standard residues are included in parenthesis. Please check the sequence information carefully. If you think it was incorrectly extracted, please check if TER cards were added correctly. PDB format requires all polymer chains should have a TER card at the end and no TER cards should be included at the end of non-polymer residues (such as ions, ligands, waters).

CHAIN A: Total 353 residues, first residue (6 ALA), last residue (356 SER)

AGSFMSVQYVQDPSRYITLQYVTEGSMVGSCAHN1NCKV1N1K1SPFN2H2QYRCRLRKL
KTNIVGINDIRAPTIEEQMDYVTQQLMSTOKQ1LQSLHDIERYL1QNYIQ1LAHNLH
ADRSV1NIUTUCQDDQDSPLAP1QDFH1ORSF1TPC1E1STP1VAPF1E1LNS1MYSK1T1
DINSV1C1AMELNSN1R1PQPSH1GLQ1NH1G1LGS1PQ1ED1NC1IN1K1MY1LS1L1PH1N1V1P1N1L1F
PHN1S1K1AL1CD1L1D1L1T1F1N1K1RIVE1G1AL1H1Y1LE1Q1Y1D1F1S1P1A1E1AP1K1F1D1M1L1D1L1P1K1L1K1L1E1F1T1S1R1Q1P1Y1R1S

If a sequence above has missing residues, correct the sequence later during deposition.
Validation Report

Click here to e-mail the session restart ID

Please review the Validation Report carefully. If there are any outstanding issues such as sequence link problem, geometry errors, chirality errors etc., please make sure you fix the problem(s) and start a fresh session with corrected file.

Please note that the geometry errors shown in the Validation Report will also appear in the processed PDB file.

For ligands, please make sure the three letter codes are correct. Both extra and missing atoms may appear in the Validation Report if incorrect ligand three letter code was used. If only missing atoms are listed in the Validation Report, either incorrect ligand three letter code was used or part of ligand is missing coordinates (see Ligand Expo).

Thank you for using PDB. The following geometrical and stereoechemical features have been calculated for your structure:

CLOSE CONTACTS

-------------

=> Close contacts in same asymmetric unit. Distances smaller than 2.2 Å. Angstroms are considered as close contacts.

<table>
<thead>
<tr>
<th>Chain Atom</th>
<th>Res. Seq</th>
<th>Chain Atom</th>
<th>Res. Seq</th>
<th>Symm. Code</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>542</td>
<td>O</td>
<td>543</td>
<td>(1, 5, 5, 5)</td>
<td>Dist = 0.69</td>
</tr>
<tr>
<td>O</td>
<td>585</td>
<td>O</td>
<td>586</td>
<td>(1, 5, 5, 5)</td>
<td>Dist = 1.00</td>
</tr>
<tr>
<td>O</td>
<td>587</td>
<td>O</td>
<td>588</td>
<td>(1, 5, 5, 5)</td>
<td>Dist = 1.00</td>
</tr>
</tbody>
</table>

=> Close contacts based on crystal symmetry. Distances smaller than 2.2 Å. Angstroms are considered as close contacts.
Please input sequence(s):

Only sequences for proteins, peptides, and nucleic acids should be entered in the boxes below.

- Please enter the sequence of each macromolecular polymer. The sequence should include all residues in the experiment including uncleaved expression tags (like his tags), as well as residues missing due to disorder.
- If a sequence applies to more than one chain, enter the sequence once and specify the chain IDs with spaces or commas in between each chain ID.
- N- and C-terminal tags can be entered in the "uncleaved N/C-terminal HIS-tag or cloning artifact" boxes. The N- and C-terminal boxes should be left blank if no tags were used or if they are already listed in the sequence.
- To indicate a modified or non-standard residue like selenomethionine, use parenthesis. (MSE). Three letter codes can be found at Ligand Expo.
- The polymer type must be selected.
- If the sequence database reference exists, please select the database name and enter the accession code (GB (GeneBank) for nucleic acids and UPN (UniProt) protein sequence).
- For Structure Genomics, if the sequence has been or will be deposited to TargetDB, please indicate the Target ID in the "Target identifiers" box. This is for structural genomics project only.
- One-letter code for polymer sequence INCLUDING THE RESIDUES MISSING IN THE COORDINATES (mandatory). Do not repeat terminal tags if already provided in the boxes above or below.

Click here to e-mail the session restart ID

Save Sequence
How to use the AUTO DEP INPUT TOOL

- **BEGIN ENTERING DATA**: For each category, a set of instructions and places to enter data will be displayed in this frame. Fill in any relevant missing information.

- **RESTART**: If you want to restart your session at a later time, record the restart ID for this session: 2010-Nov-09.21.16.59.5209.Session.11011 (Click here to e-mail the session restart ID). This ID is also displayed in the title bar of your browser. When you are ready to continue, enter this code in the appropriate box at the bottom of the ADIT HOME page.

- **HELP**: An explanation for any item can be obtained by selecting the Help button within the table. This information will appear in the bottom frame. Pressing the Help button in the top frame will display these instructions.

- **EXAMPLE**: An example for any item can be obtained by selecting the Example button within the table. This information will appear in the bottom frame.

- **SAVING YOUR DATA**: You must press the Save button in each category after entering data. Switching to another category without saving your data will erase the information that you have entered.

This is the HELP frame

This frame is used to display dictionary descriptions, examples, tables and diagnostic information.

For Data Items:

- Press the HELP button for advice about how to enter data in the item.
- Press the EXAMPLE button in the item frame to view examples of the item.
Enter the crystallographic cell parameters for this structure.

**Unit Cell**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length a</td>
<td>92.50</td>
</tr>
<tr>
<td>Length b</td>
<td>92.50</td>
</tr>
<tr>
<td>Length c</td>
<td>108.0X</td>
</tr>
<tr>
<td>Angle alpha</td>
<td>90.00</td>
</tr>
<tr>
<td>Angle beta</td>
<td>90.00</td>
</tr>
<tr>
<td>Angle gamma</td>
<td>90.00</td>
</tr>
</tbody>
</table>

**Examples of Angle gamma**

(mmCIF item_cell.angle.gamma)

- Example 1:
  - 90.00
- Example 2:
  - 72.96
Required Information

Deposition
- Contact Authors
- Structural Genomics
- Release Status
- Title
- Related Entries

Authors
- Entry Authors
- Contact Authors
- Citation

Chemical/Biological Features
- Molecule Names
- Molecule Details
- Sequence
- Genetically Manipulated Source
- Natural Source
- Synthetic Source

Structure Features
- Keywords
- Biological Assembly
Quantity and Quality of PDB data
Differences of views

Results (or PDB ID)

PDB files

Source

Structural Biologists

Bioinformatics
Data Quantity

- Fill in as many data items as possible (various small tools available)
- Notify us when the paper gets published so that we can release the entry
Data Quality

It is up to depositors (you!!)…

but we have done about our part of it.
Common troubles observed...

- Wrong or missing data items
  ("to be published," species, R-factors etc)
- Redundant definition
  ("human" or "homo sapience")
- Ligand structure and definition
**mmCIF-specific troubles**

- Lack of validation program of contents
  
  \[_\text{phasing.method} \ '\text{MAGIC}'\]

- (Inconsistent) redundancy of data items

```
mmCIF → non-redundant XML (PDBj-ML) → mmCIF
```
PDB Remediation Project

Thorough check of the PDB contents by all wwPDB members.
New dataset was released on 1 August 2007.

What went wrong?

Paper

PDB
Related Database/Services
at PDBj
PDBj Mine

---

**Home**

Data Deposition >>
- ADIT: PDB Deposition
- ADIT-NMR

Search >>
- Search PDB (Mine/xPSSS)
- Latest Released Search
- Sequence-Navigator
- Structure-Navigator

Search and Software >>
- SeSaw
- Ligand Binding Sites (GIRAFF)
- EM Navigator
- Search NMR Data (BMRB)
- Status Search

Service and Software >>
- jV: Graphic Viewer
- Protein Globe
- ASH
- MAFFTash
- SEALA

Structure Prediction >>
- CRMPRED

---

**Deposition**

PDBj (Protein Data Bank Japan) maintains a centralized archive of macromolecular structures and provides integrated tools, in collaboration with the RCSB, the BMRB in USA and the PDBs in EU. PDBj is supported by JST-BIRD.

**Search**

Search PDB

Search NMR Data

---

**What's new**

6-Oct-2010

Effective December 6, 2010, deposition of chemical shift data will be mandatory when submitting NMR entries to the PDB. (more...)

---

**69162 entries available on 10 Nov., 2010**

00:00(UTC) / 09:00(JST)
PDBj (Protein Data Bank Japan) maintains a centralized archive of macromolecular structures and provides integrated tools, in collaboration with the RCSB, the BMRB in USA and the PDBe in EU. PDBj is supported by JST-BIRD.

**Summary [12as]**

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>12as_sequence information (FASTA format)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptor</td>
<td>ASPARAGINE SYNTHETASE, L-ASPARAGINE, ADENOSINE MONOPHOSPHATE</td>
</tr>
<tr>
<td>Title</td>
<td>ASPARAGINE SYNTHETASE MUTANT C51A, C315A COMPLEXED WITH L-ASPARAGINE AND AMP</td>
</tr>
<tr>
<td>Functional Keywords</td>
<td>LIGASE, ASPARAGINE SYNTHETASE, NITROGEN FIXATION</td>
</tr>
<tr>
<td>Biological source</td>
<td>Escherichia coli K12</td>
</tr>
<tr>
<td>Cellular location</td>
<td>[UNP-P00963] Cytoplasm</td>
</tr>
<tr>
<td>Total number of polymer chains</td>
<td>2</td>
</tr>
<tr>
<td>Total molecular weight</td>
<td>74226 (the details in Structural Details Page)</td>
</tr>
<tr>
<td>Authors</td>
<td>Nakatsu, T., Kato, H., Oda, J. (deposition date: 1997-12-02, release date: 1998-12-30)</td>
</tr>
</tbody>
</table>

Structure Viewers

**jV3 / Jmol**

(jV3 and Jmol require Java (TM) Plug-in 1.5 or later.)
Graphic viewer: jV version 3.8
Download from http://www.pdbj.org/jV/

jV version 3 (formerly known as PDBjViewer) is a program to display molecular graphics of proteins and nucleic acids. jV supports the following features:

- jV can read and display PDBML files, the canonical XML format for the Protein Data Bank.
- Of course, jV can read and display the traditional PDB format files, too.
- RasMol-like usability.
- jV can process more than one molecules.
- jV can display polygons specified by XML. (XML Schema for polygons is available.)
- Multiple polygons can be processed simultaneously, and be superimposed onto molecular images.
- Animation can be realized.
- jV runs on the Java Runtime Environment (JRE) so that it can work as a stand-alone application as well as an applet.
- The graphics of jV is based on OpenGL (JOGL), thereby producing fairly beautiful pictures.

Download
Development of other Databases & Services

Protein Folds Browser, Protein Globe (Kinjo & Standley)

Ligand Binding Site Search, GIRAF (Kinjo)

Electron Microscopy Navigator, EM-Navi (Suzuki)

Protein Molecular Surface Database, eF-site (Kinoshita & Nakamura)

Search for Similar Surface, eF-seek (Kinoshita & Nakamura)

Protein Dynamics Database, ProMode (Wako & Endo)
Development of other Databases & Services

Homolog protein search, Sequence Navigator (Standley)

Similar fold search, Structure Navigator (Standley & Toh)

Alignment of Sequence and Structures, MAFFTash (Kato, Toh & Standley)

Encyclopedia of Protein Structures, eProtS (Kinjyo, Kudo, & Ito)

Molecule of the Month, MoM (Goodsell & Kudo)

Function Annotation from Folds and Sequences, SeSAW (Standley)