Deposition and Validation using RCSB PDB Tools

Or, how to make your life (and mine) easier

Kyle Burkhardt, Data Annotation Leader

RCSB PDB at Rutgers University

www.pdb.org

deposit@rcsb.rutgers.edu

RCSB PDB
PROTEIN DATA BANK

Why are you here?

- Learn how to deposit data quickly, easily, accurately, and efficiently
- Learn about RCSB PDB tools
  - pdb_extract, Validation Server, Ligand Depot, ADIT
- Learn how the RCSB PDB annotates structures

Why do you deposit your structural data to the PDB?

- "Compulsory" reasons
  - Primary citation journal policies requires it
  - Funding agency requires it
- "Voluntary" reasons
  - For safe-keeping of structural data
  - For the benefit of the entire scientific community

When do you deposit?

- Immediately after structure determination
- Just prior to or after submission of manuscript
- After the manuscript has been accepted – urgent request for PDB ID
- Just before the researcher is leaving the lab
- Several years after the initial data collection

What do you deposit?

- The coordinates
- The structure factor file(s)
- and more …
  - Information that only you can provide
  - Information that you should complete and verify
    - about the molecule(s) or complex
    - about the crystallization and data collection
  - Information that can be extracted from log files of crystallographic applications.

How and Where do you deposit?

- Using the ADIT tool
  - http://deposit.pdb.org/adit/ (RCSB-PDB) or
  - http://pdbdep.protein.osaka-u.ac.jp/adit/ (PDBj).
- Using AutoDep
  - http://autodep.ebi.ac.uk/ (MSD/EBI).

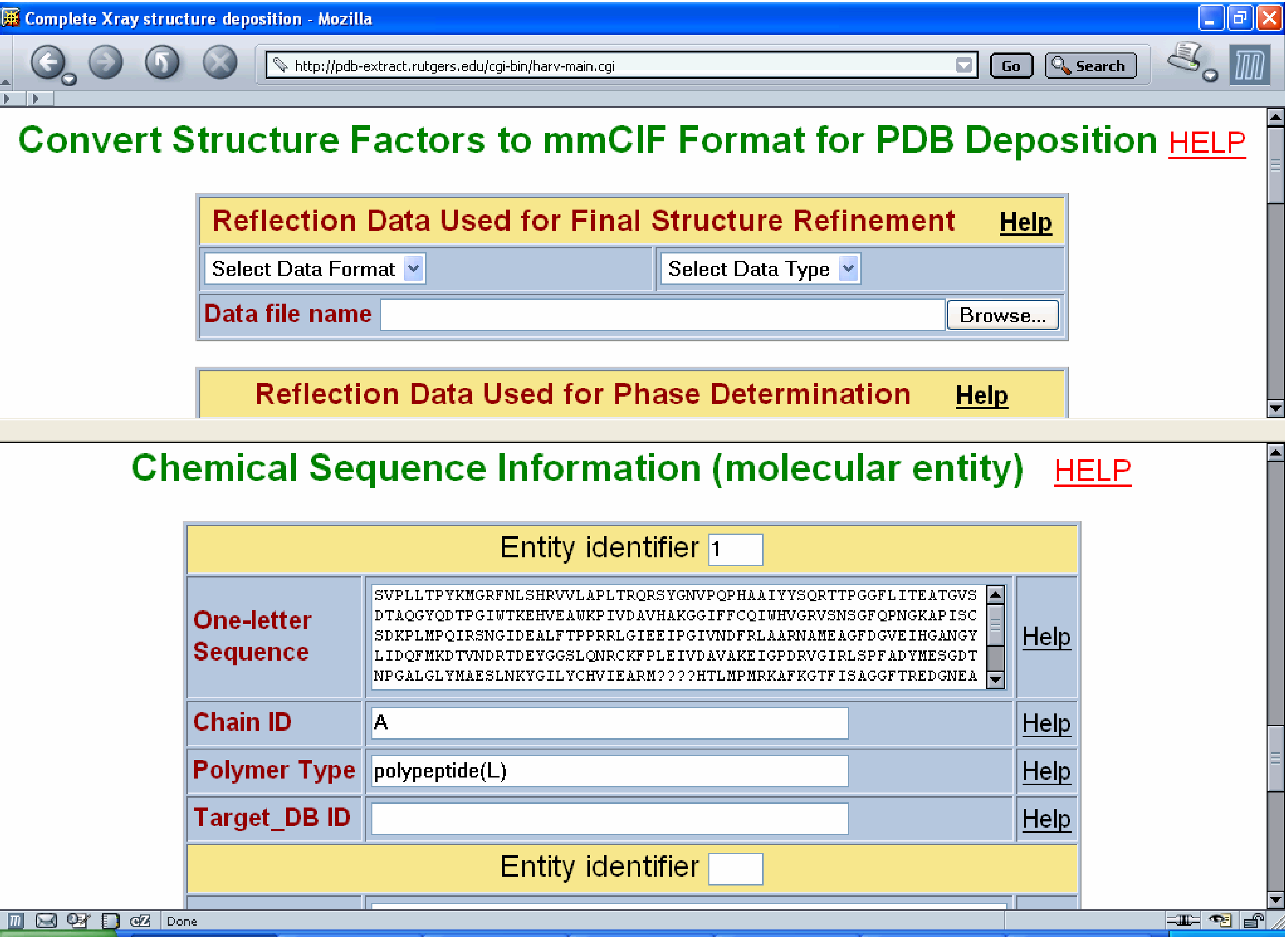
5 Easy Steps for Fast, Accurate, and Complete Data Deposition at the RCSB PDB

1. **Use pdb_extract**
2. **Validate your entry**
3. **Verify sequence**
4. **Use Ligand Depot**
5. **Deposit with ADIT**

This is an iterative process

1. Use PDB EXTRACT

- RCSB pdb_extract
  - Extracts data from crystal structure determination programs
  - Fills in many fields automatically
  - Template file for multiple depositions
    - Fill in protein name, citation, status, source, author info once and use template multiple times
  - Generates a complete data file ready for deposition



Complete Xray structure deposition - Mozilla

http://pdb-extract.rutgers.edu/cgi-bin/harv-main.cgi

Convert Structure Factors to mmCIF Format for PDB Deposition HELP

Reflection Data Used for Final Structure Refinement Help

Select Data Format | Select Data Type

Data file name: [ ] Browse...

Reflection Data Used for Phase Determination Help

Chemical Sequence Information (molecular entity) HELP

Entity identifier: 1

One-letter Sequence	SVPLLTPYKMGRFNLSHRVVLAPLTRQRSYGNVPQPHAAIYYSQRTTPGGFLITEATGVS DTAQGYQDTPGIWTKEHVEAWKPIVDAVHAKGGIFFCQIWHVGRVSNSGFQPNGKAPISC SDKPLMPQIRSNGIDEALFTPPRRLGIEEIPGIVNDFRLAARNAMEAGFDGVEIHGANGY LIDQFMKDTVNDRTDEYGGSLQNRCKFPLEIVDAVAKEIGPDRVGIRLSPFADYMESGDT NPGALGLYMAESLNKYGILYCHVIEARM????HTLMPMRKAFKGTFISAGGFTREDGNEA	Help
Chain ID	A	Help
Polymer Type	polypeptide(L)	Help
Target_DB ID		Help

Entity identifier:

PDB EXTRACT Availability

- CCP4 package (CCP4i interface)
  - Script and command line
- Desktop (script and command line)
  - ***sw-tools.pdb.org/apps/PDB_EXTRACT***
- Web-based
  - ***pdb-extract.rutgers.edu***
- Tutorial
  - ***pdb-extract.rutgers.edu/tutorial.html***

2. Validate Your Entry

ADIT! Validation Server

- RCSB PDB Validation Server
  - Reads mmCIF file from pdb_extract
  - Reads PDB or mmCIF files from refinement programs
  - Reads structure factor file in mmCIF format
- Steps in Validation
  1. Precheck coordinate and experimental data files
  2. Produce validation report

Validation Reports Contain:

- Close contacts
- Bond and angle deviations
- Chirality errors
- Sequence/coordinate (mis)alignment
- Missing and extra atoms or residues
- Distant waters
- NUCheck[1], PROCHECK[2], SFCHECK[3] , MolProbity[4] Reports

[1] Feng Z, Westbrook J, Berman HM.(1998) NUCheck: Rutgers University, New Brunswick, NJ. Report No.: NDB-407.

[2] Laskowski, R.A., McArthur, M.W., Moss, D.S., et al. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst*. 26:283-291.

[3] 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.

[4] Lovell SC, Davis IW, Arendall III WB, de Bakker, PIW, Word JM,Prisant MG, Richardson JS, Richardson DC (2003). Structure Validation by C-alpha Geometry: phi, psi and C-beta Deviation. *PROTEINS: Structure, Function, and Genetics*, **50**, 437-450.

Validation Availability

- Desktop
  - ***sw-tools.pdb.org/apps/VAL/***
- Web-based
  - ***pdb.rutgers.edu/validate/***
- pdb_extract
  - Command line option
- ADIT
  - Desktop and Web
- Tutorial
  - ***deposit.pdb.org/validate/docs/tutorial.html***

3. Verify Sequence

- Input the complete deposition sequence

(e.g. BLAST ***www.ncbi.nih.gov/BLAST***[1])

  - Include
    - residues missing due to lack of electron density
    - cloning artifacts and HIS tags that were not cleaved
    - mutations or substitutions
- Output compares the deposition sequence to sequence database references.
- Check sequence database correspondence

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol*. 215:403-410

Sequence Discrepancies

- Alanine or glycine mismatches
- GLU/GLN or ASP/ASN mismatches
- Intended mutation
- Deletion or insertion
- Unobserved gap
- Real or unexpected difference
  - “We’re right and they’re wrong”

Sample BLAST output

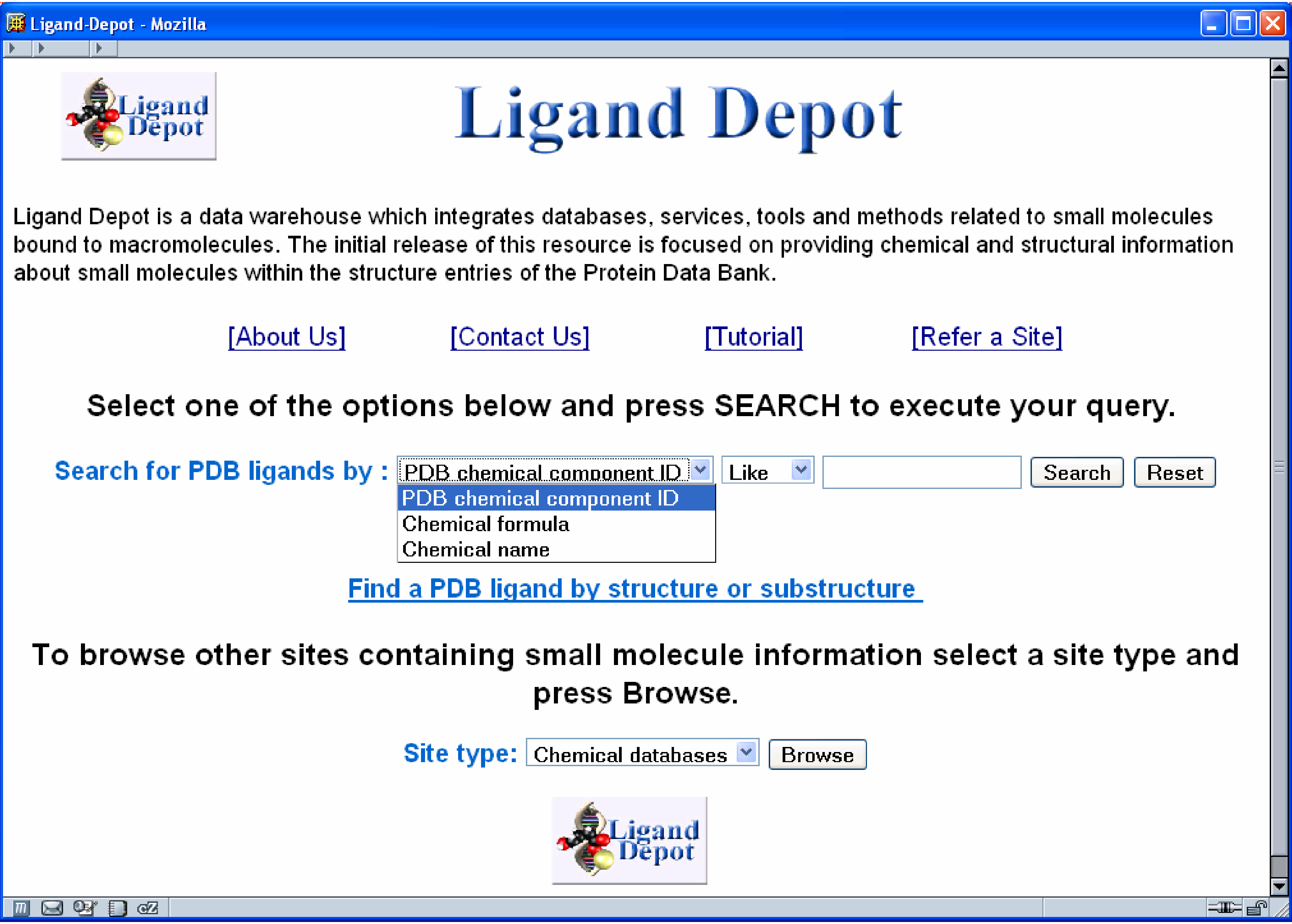
```
>gi|126605|sp|P00720|LYCV_BPT4 Lysozyme (Lysis protein) (Muramidase)
   (Endolysin)
          Length = 164

 Score =  189 bits (440), Expect = 2e-48
 Identities = 65/80 (81%), Positives = 67/80 (83%), Gaps = 6/80 (7%)

Query: 1  MNIFEMLRIDQGLAAAAAANTEGYYTIGIGHLLT------AAKSELDKAIGRNTNGVITK 54
          MNIFEMLRID+GL      +TEGYYTIGIGHLLT      AAKSELDKAIGRN NGVITK
Sbjct: 1  MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITK 60
```

4. Use

- RCSB PDB Ligand Depot
  - Use to find code for existing ligands
  - Searching by many attributes
  - New ligands
    - E-mail chemical diagram (with bond order), IUPAC name, synonyms, and formula to ***deposit@rcsb.rutgers.edu***
    - Choose your three letter code for new ligands
- Access
  - ***ligand-depot.rutgers.edu***



Ligand Depot

Ligand Depot is a data warehouse which integrates databases, services, tools and methods related to small molecules bound to macromolecules. The initial release of this resource is focused on providing chemical and structural information about small molecules within the structure entries of the Protein Data Bank.

[About Us] [Contact Us] [Tutorial] [Refer a Site]

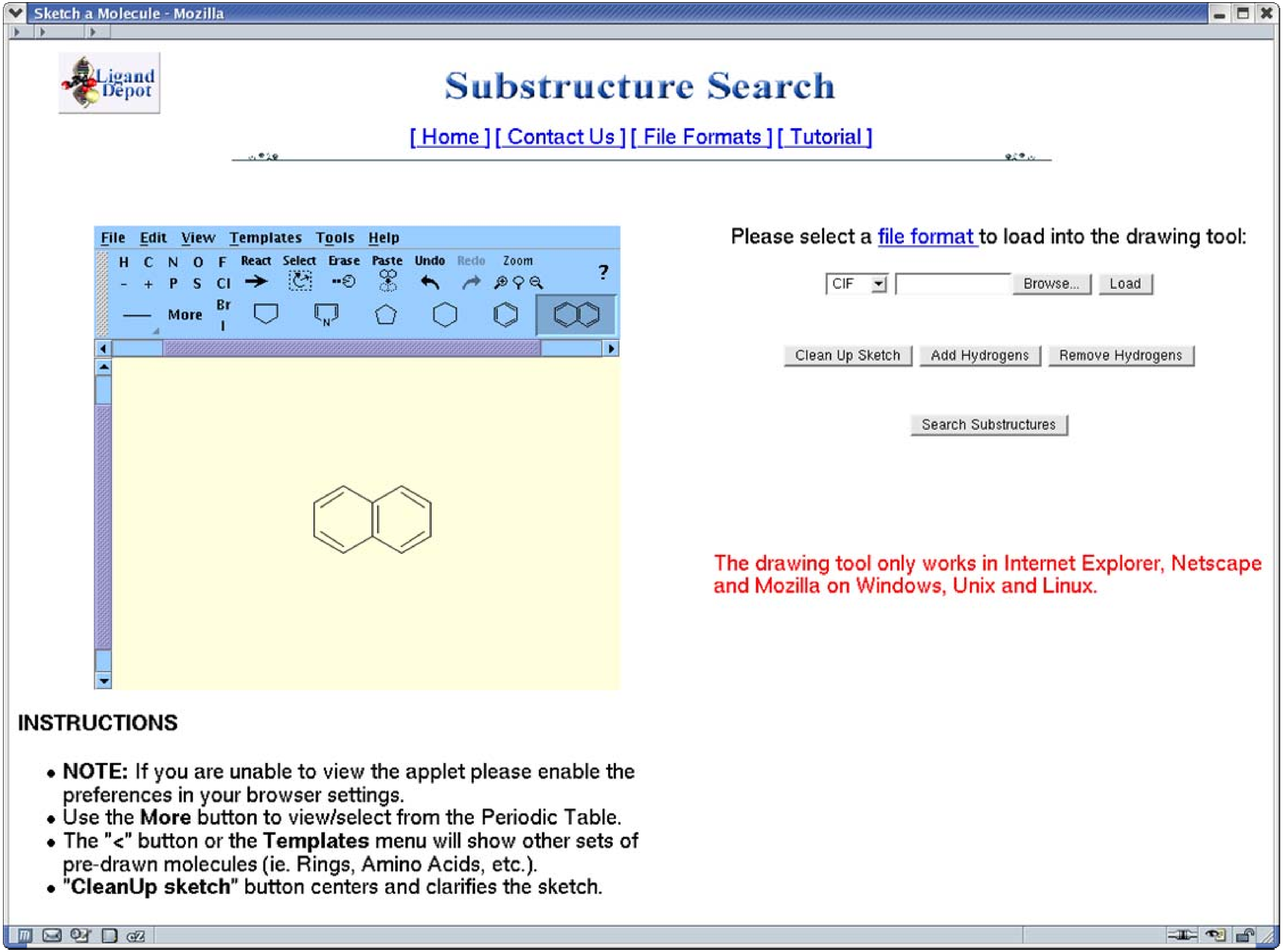
Select one of the options below and press SEARCH to execute your query.

Search for PDB ligands by : PDB chemical component ID | Like

Find a PDB ligand by structure or substructure

To browse other sites containing small molecule information select a site type and press Browse.

Site type: Chemical databases



Substructure Search

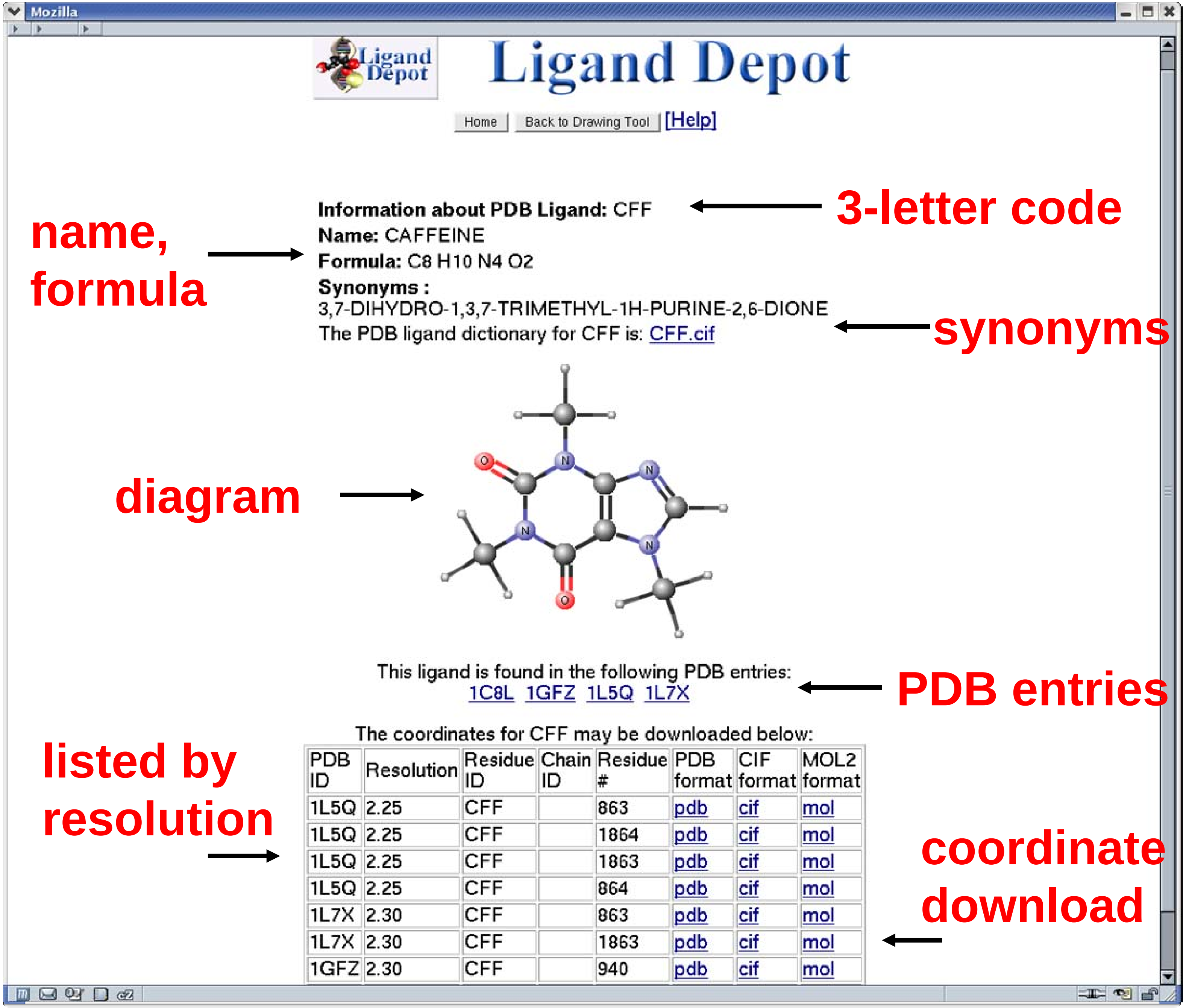
[Home] [Contact Us] [File Formats] [Tutorial]

Please select a file format to load into the drawing tool:

The drawing tool only works in Internet Explorer, Netscape and Mozilla on Windows, Unix and Linux.

INSTRUCTIONS

- **NOTE:** If you are unable to view the applet please enable the preferences in your browser settings.
- Use the **More** button to view/select from the Periodic Table.
- The "**<**" button or the **Templates** menu will show other sets of pre-drawn molecules (ie. Rings, Amino Acids, etc.).
- "**CleanUp sketch**" button centers and clarifies the sketch.



Ligand Depot

Information about PDB Ligand: CFF

Name: CAFFEINE

Formula: C8 H10 N4 O2

Synonyms :
3,7-DIHYDRO-1,3,7-TRIMETHYL-1H-PURINE-2,6-DIONE

The PDB ligand dictionary for CFF is: CFF.cif

This ligand is found in the following PDB entries:
1C8L 1GFZ 1L5Q 1L7X

The coordinates for CFF may be downloaded below:

PDB ID	Resolution	Residue ID	Chain ID	Residue #	PDB format	CIF format	MOL2 format
1L5Q	2.25	CFF		863	pdb	cif	mol
1L5Q	2.25	CFF		1864	pdb	cif	mol
1L5Q	2.25	CFF		1863	pdb	cif	mol
1L5Q	2.25	CFF		864	pdb	cif	mol
1L7X	2.30	CFF		863	pdb	cif	mol
1L7X	2.30	CFF		1863	pdb	cif	mol
1GFZ	2.30	CFF		940	pdb	cif	mol

3-letter code

name, formula

synonyms

diagram

PDB entries

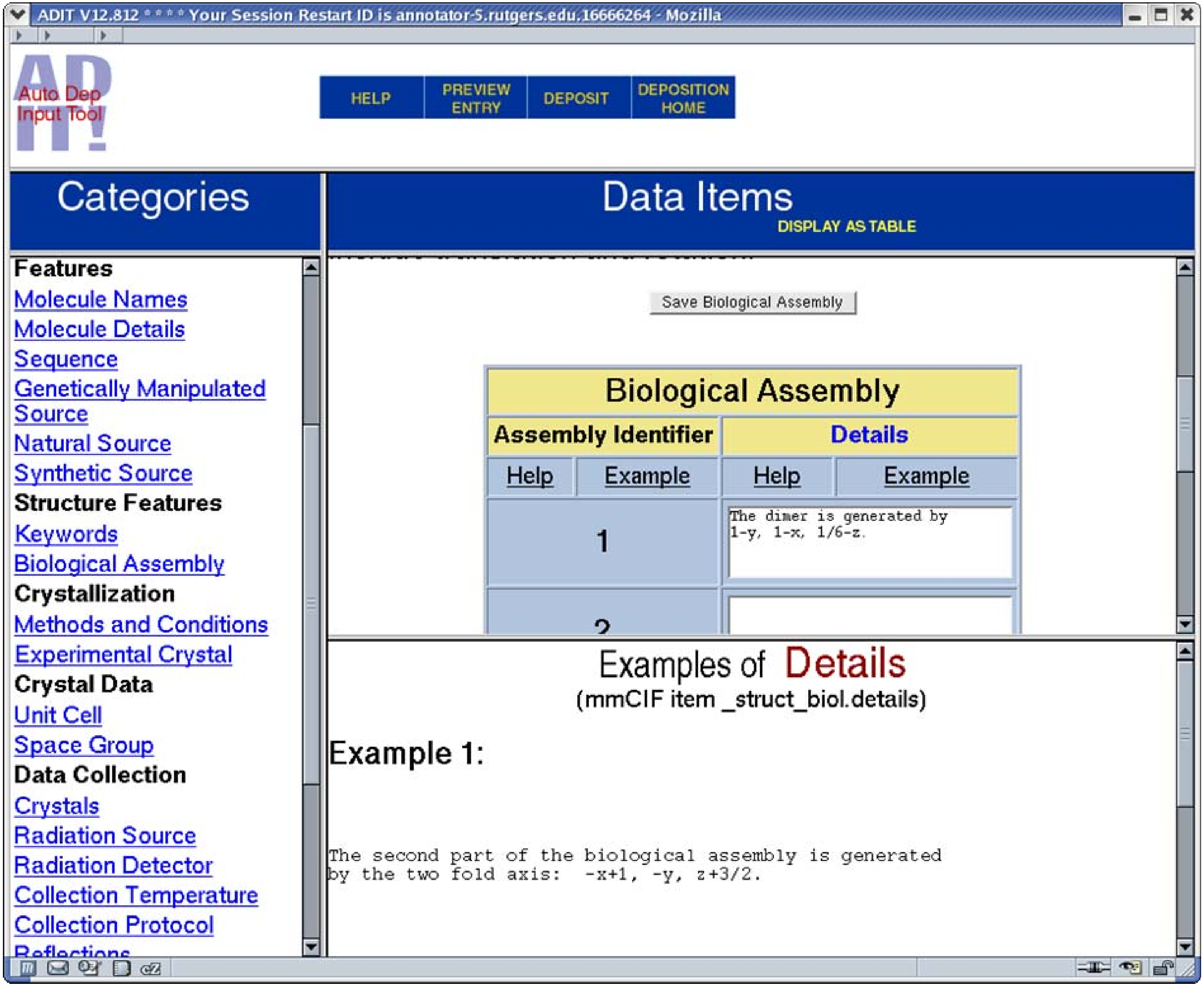
listed by resolution

coordinate download

5. Deposit with ADIT! Auto Dep Input Tool

- Web-based ADIT ***(deposit.pdb.org/adit/)***
  - Load file (coordinates and sfs)
    - Input missing information
  - Deposit
- Desktop ADIT ***(sw-tools.pdb.org/apps/ADIT)***
  - Load file (coordinates and sfs), add missing information, validate and save
  - Deposit
    - Load in Web-based ADIT and deposit

Tutorial ***deposit.pdb.org/adit/docs/tutorial.html***



Biological Assembly			
Assembly Identifier		**Details**	
Help	Example	Help	Example
1		The dimer is generated by 1-y, 1-x, 1/6-z.	
2			

Examples of Details
(mmCIF item _struct_biol.details)

Example 1:

```
The second part of the biological assembly is generated
by the two fold axis:  -x+1, -y, z+3/2.
```

Important Points to Consider

- Title
- Sequence (including mutations)
- Protein name
- Biological unit
- Ligands
- Visual inspection of the entry
- Unusual situations

Don’t be shy. Talk is good.
Tell us the whole story right away.

I just deposited. Now what happens?

- **What is annotation?**
  - *"A note added by way of comment or explanation"*
- **When do we annotate?**
  - *All the time! You never stop depositing*
- **Where do we annotate?**
  - *RCSB-PDB @ Rutgers and Prague*
  - *Who else annotates? MSD/EBI, PDBj*
- **Why do we annotate?**
  - *Annotators are here to help you represent your data in the best possible way*
- **How do we annotate?**
  - *We use the same tools we want you to use*

What Do Annotators Do?

- **Annotators check everything**
  - Check entry for self-consistency
  - Check title
  - Check citation references with PubMed (http://pubmed.gov/)
  - Correct format errors in data and coordinates
  - Check sequence
  - Add sequence database reference
  - Add protein name and synonyms
  - Check source
  - Check ligand nomenclature
  - Add biological unit information
  - Visually check entry
  - Generate validation reports

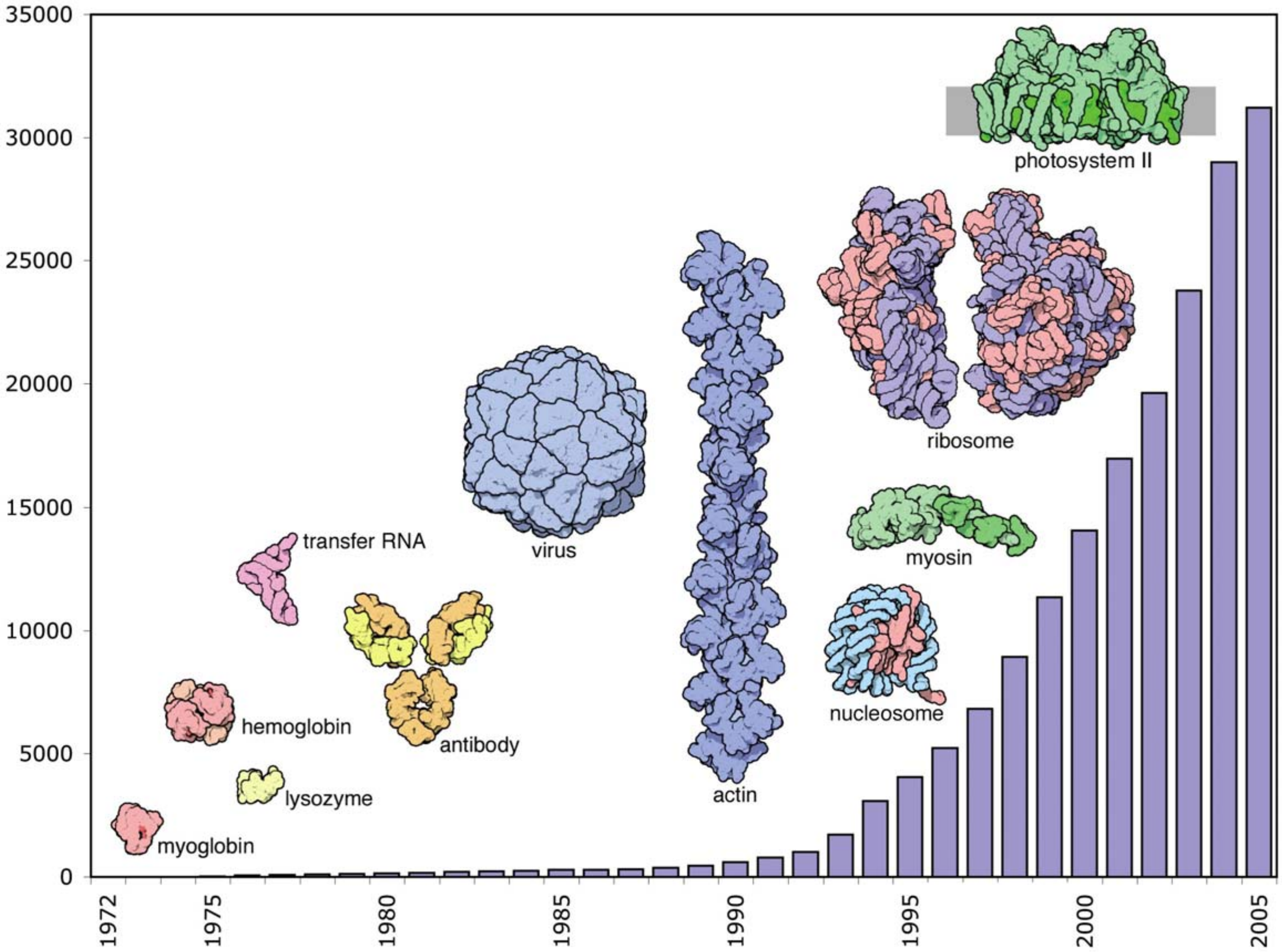
After Initial Annotation

- Correspond with you
- Update entries
  - Corrections, new coordinate sets
- Release entries
- How long does the entire process take?

  It's dependent on…
  - Number and type of corrections
  - New coordinate set(s)

- Annotation is like a box of chocolates…

Growth of the PDB archive



Release Information

- **Release options**
  - Pre-release of sequence
  - Coordinate release
    - Release immediately
    - Hold until publication (HPUB)
    - Hold until a particular date
    - HPUB and HOLD limit
      - not more than 1 year after deposition

- **It’s ok to release a structure without a citation**

How Do We Find Citations?

- Some journals
- PDB users
- Weekly PubMed searches
- You tell us (please tell us!)

How to make my life more difficult

1. Deposit before you finish your refinement
2. Don’t use any software tools
3. Don’t follow deposition procedures
4. Deposit the day you leave your company or postdoc position
5. Deposit the day the journal needs the PDB ID
6. Don’t tell us the whole story until later
7. Provide only one contact author
8. Don’t tell us when the structure has been published
9. Don’t acknowledge our e-mail

It makes your life more difficult too…

Please do the following…

- Give yourself time to deposit
- Use pdb_extract
- **Validate** (check your data) before deposition
- Verify the sequence
- Use Ligand Depot
- Communicate with us

Annotator attitudes are influenced by you

If you don’t validate → If you do validate

Why Should You Do What I Say?

- Create a more complete deposition with less manual input
- Minimize mistakes
  - Check (and recheck)
  - Give yourself time to deposit
- Save time (for you and us)
- Help us help you!

The RCSB PDB annotation staff thanks you!

Monica Sundd, Shuchismita Dutta, Jasmine Young, Kyle Burkhardt,
Jeramia Ory, Shri Jain, Massy Rajabzadeh, Irina Persikova
Not pictured: Bohdan Schneider

Acknowledgements

- The RCSB Protein Data Bank (PDB) is operated by
  - Rutgers, The State University of New Jersey
  - San Diego Supercomputer Center at the University of California, San Diego

- The RCSB PDB is supported by funds from
  - National Science Foundation (NSF)
  - National Institute of General Medical Sciences (NIGMS)
  - Office of Science, Department of Energy (DOE)
  - National Library of Medicine (NLM)
  - National Cancer Institute (NCI)
  - National Center for Research Resources (NCRR)
  - National Institute of Biomedical Imaging and Bioengineering (NIBIB)
  - National Institute of Neurological Disorders and Stroke (NINDS)

- The worldwide PDB (wwPDB) is a collaboration between
  - RCSB
  - MSD/EBI
  - PDBj

RCSB PDB Data Deposition Services

- **pdb_extract**
  - **Web- http://pdb-extract.rutgers.edu/**
  - **Standalone - http://*sw-tools.pdb.org/apps/PDB_EXTRACT***
- **Validation Server**
  - **Web - http://deposit.pdb.org/validate/**
  - **Standalone - http://*sw-tools.pdb.org/apps/VAL/***
- **ADIT**
  - **Web – http://deposit.pdb.org/adit/**
  - **Standalone - http://*sw-tools.pdb.org/apps/ADIT***
- **Ligand Depot - http://ligand-depot.rutgers.edu/**
- **Overview and tutorials for all RCSB PDB data deposition services – http://deposit.pdb.org**