Querying the PDBj Mine2 relational database

Akira R. Kinjo

Institute for Protein Research, Osaka University
&
Protein Data Bank Japan (PDBj)

wwPDB File Formats

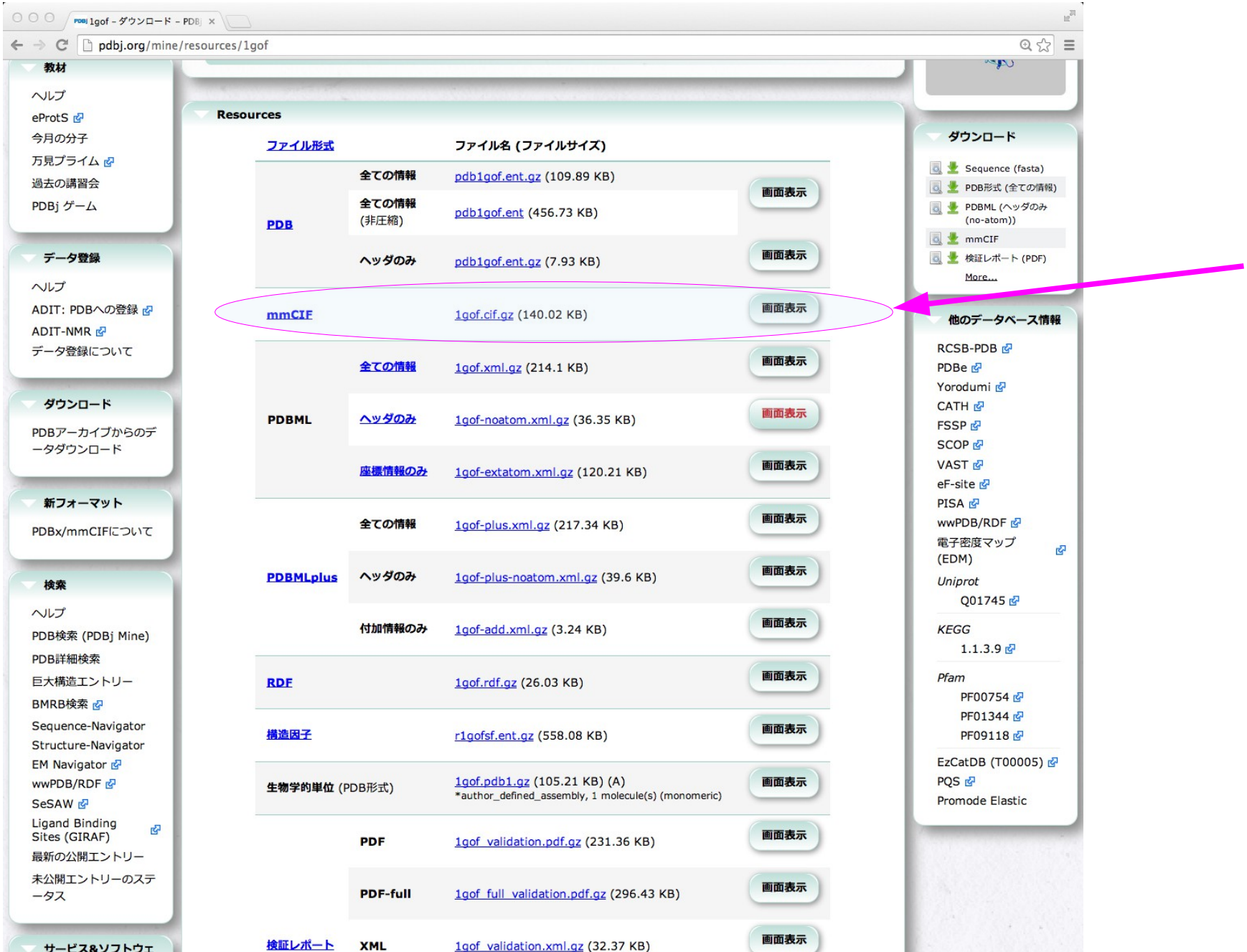
File formats

- mmCIF
  - The canonical format of the wwPDB.
  - Ver. 5 released this year!
- PDBML
  - “direct translation” of mmCIF into XML.
- (Legacy) PDB format
  - *NOT RECOMMENDED!*
- PDB/RDF
  - Translation of PDBML into RDF/XML (the standard format for the Semantic Web).

Other data

- Validation Report
- Structure factor (for crystal structures)
- NMR distance restraints

mmCIF: an example



mmCIF basics

- Data are divided into “categories”.
  - *_category.item*
  - e.g., _entry.id “entry” is the category name, “id” is an item of the “entry” category.
  - “_entry.id 1GOF” The value of “id” item of “entry” category is “1GOF”.
- Two ways of presenting data.
  - key-value: if only one value exists for an item.
  - loop: if more than one item exists for an item.

More about mmCIF

- Context-free grammar (STAR [Self-def ning Text Archive and Retrieval])
- All the categories and items are def ned in the PDBx/mmCIF dictionary.
- For details, see http://mmcif.wwpdb.org/

A closer look at mmCIF

```
data_1GOF
#
_entry.id   1GOF
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    4.007
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
_database_2.database_id     PDB
_database_2.database_code   1GOF
#
….
```

This tag starts a data (“datablock” is the unit of data [entry])

entry ID (PDB ID) is “1GOF”.

Provenance information is also included.

Example of key-value pairs.

```
_cell.entry_id                  1GOF
_cell.length_a                  98.000
_cell.length_b                  89.400
_cell.length_c                  86.700
_cell.angle_alpha               90.00
_cell.angle_beta                117.80
_cell.angle_gamma               90.00
_cell.Z_PDB                     4
_cell.pdbx_unique_axis          ?
#
```

The last “#” is a convention to indicate the end of a category.

Example of “loop” structure

```
loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
_entity.details
_entity.pdbx_mutation
_entity.pdbx_fragment
_entity.pdbx_ec
1 polymer     man 'GALACTOSE OXIDASE' 68579.250 1   ? ? ? 1.1.3.9
2 non-polymer syn 'COPPER (II) ION'   63.546    1   ? ? ? ?
3 non-polymer syn 'SODIUM ION'        22.990    1   ? ? ? ?
4 non-polymer syn 'ACETIC ACID'       60.052    2   ? ? ? ?
5 water       nat water               18.015    316 ? ? ? ?
#
```

Start of a loop

A list of items.
“One item per line” is just a convention.

- Each item is whitespace-delimited.
- In the same order as the item list (above).
- Use quotes (') for data with whitespace.

The last “#” is a convention to indicate the end of a loop.

Atomic coordinates in mmCIF

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . ALA A 1 1   ? 38.840  0.236   1.012   1.00 34.65 ? ? ? ? ? ? 1    ALA A N   1
ATOM   2    C  CA  . ALA A 1 1   ? 38.356  -0.999  0.357   1.00 42.26 ? ? ? ? ? ? 1    ALA A CA  1
ATOM   3    C  C   . ALA A 1 1   ? 37.098  -1.547  1.056   1.00 41.25 ? ? ? ? ? ? 1    ALA A C   1
ATOM   4    O  O   . ALA A 1 1   ? 36.619  -0.946  2.028   1.00 29.44 ? ? ? ? ? ? 1    ALA A O   1
ATOM   5    C  CB  . ALA A 1 1   ? 39.398  -2.114  0.379   1.00 40.70 ? ? ? ? ? ? 1    ALA A CB  1
ATOM   6    N  N   . SER A 1 2   ? 36.610  -2.666  0.495   1.00 32.67 ? ? ? ? ? ? 2    SER A N   1
ATOM   7    C  CA  . SER A 1 2   ? 35.411  -3.244  1.202   1.00 34.90 ? ? ? ? ? ? 2    SER A CA  1
ATOM   8    C  C   . SER A 1 2   ? 35.683  -4.740  1.081   1.00 38.30 ? ? ? ? ? ? 2    SER A C   1
ATOM   9    O  O   . SER A 1 2   ? 36.827  -5.147  0.747   1.00 28.59 ? ? ? ? ? ? 2    SER A O   1
ATOM   10   C  CB  . SER A 1 2   ? 34.063  -2.660  0.823   1.00 24.49 ? ? ? ? ? ? 2    SER A CB  1
ATOM   11   O  OG  . SER A 1 2   ? 33.031  -3.308  1.686   1.00 20.37 ? ? ? ? ? ? 2    SER A OG  1
```

Main category groups

- _entity (Info about molecules)
  - entity, entity_poly, pdbx_entity_nonpoly, ...
- _atom (Info about atoms)
  - atom_site
- _struct (structural info)
  - struct, struct_conf, struct_sheet, struct_conn, pdbx_struct_assembly, ...
- _chem_comp (chemical components)
  - chem_comp
- _citation (literature info)
  - citation, citation_author, ...

Relation between categories

```
loop_
_struct_asym.id
_struct_asym.pdbx_blank_PDB_chainid_flag
_struct_asym.pdbx_modified
_struct_asym.entity_id
_struct_asym.details
A N N 1 ?
B N N 2 ?
C N N 3 ?
D N N 4 ?
E N N 4 ?
F N N 5 ?
#
```

“child” category

Parent-child relations are also def ned in the dictionary.

“parent” category

```
loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
_entity.details
_entity.pdbx_mutation
_entity.pdbx_fragment
_entity.pdbx_ec
1 polymer     man 'GALACTOSE OXIDASE' 68579.250 1   ? ? ? 1.1.3.9
2 non-polymer syn 'COPPER (II) ION'   63.546    1   ? ? ? ?
3 non-polymer syn 'SODIUM ION'        22.990    1   ? ? ? ?
4 non-polymer syn 'ACETIC ACID'       60.052    2   ? ? ? ?
5 water       nat water               18.015    316 ? ? ? ?
#
```

“label” and “auth”

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N N      . ALA A 1 4   ? 11.751 37.846 29.016  1.00 44.65 ? ? ? ? ? ? 4   ALA A N    1
ATOM   2    C CA     . ALA A 1 4   ? 12.501 39.048 28.539  1.00 30.68 ? ? ? ? ? ? 4   ALA A CA   1
ATOM   3    C C      . ALA A 1 4   ? 13.740 38.628 27.754  1.00 24.74 ? ? ? ? ? ? 4   ALA A C    1
ATOM   4    O O      . ALA A 1 4   ? 14.207 37.495 27.890  1.00 25.59 ? ? ? ? ? ? 4   ALA A O    1
ATOM   5    C CB     . ALA A 1 4   ? 12.902 39.919 29.730  1.00 16.77 ? ? ? ? ? ? 4   ALA A CB   1
ATOM   6    N N      . TYR A 1 5   ? 14.235 39.531 26.906  1.00 19.29 ? ? ? ? ? ? 5   TYR A N    1
ATOM   7    C CA     . TYR A 1 5   ? 15.552 39.410 26.282  1.00 8.51  ? ? ? ? ? ? 5   TYR A CA   1
ATOM   8    C C      . TYR A 1 5   ? 16.616 38.913 27.263  1.00 6.11  ? ? ? ? ? ? 5   TYR A C    1
ATOM   9    O O      . TYR A 1 5   ? 17.187 37.844 27.068  1.00 17.99 ? ? ? ? ? ? 5   TYR A O    1
ATOM   10   C CB     . TYR A 1 5   ? 15.988 40.762 25.702  1.00 2.00  ? ? ? ? ? ? 5   TYR A CB   1
…
```

“label_...”: labels assigned by wwPDB.

“auth_...”: labels assigned by the authors.

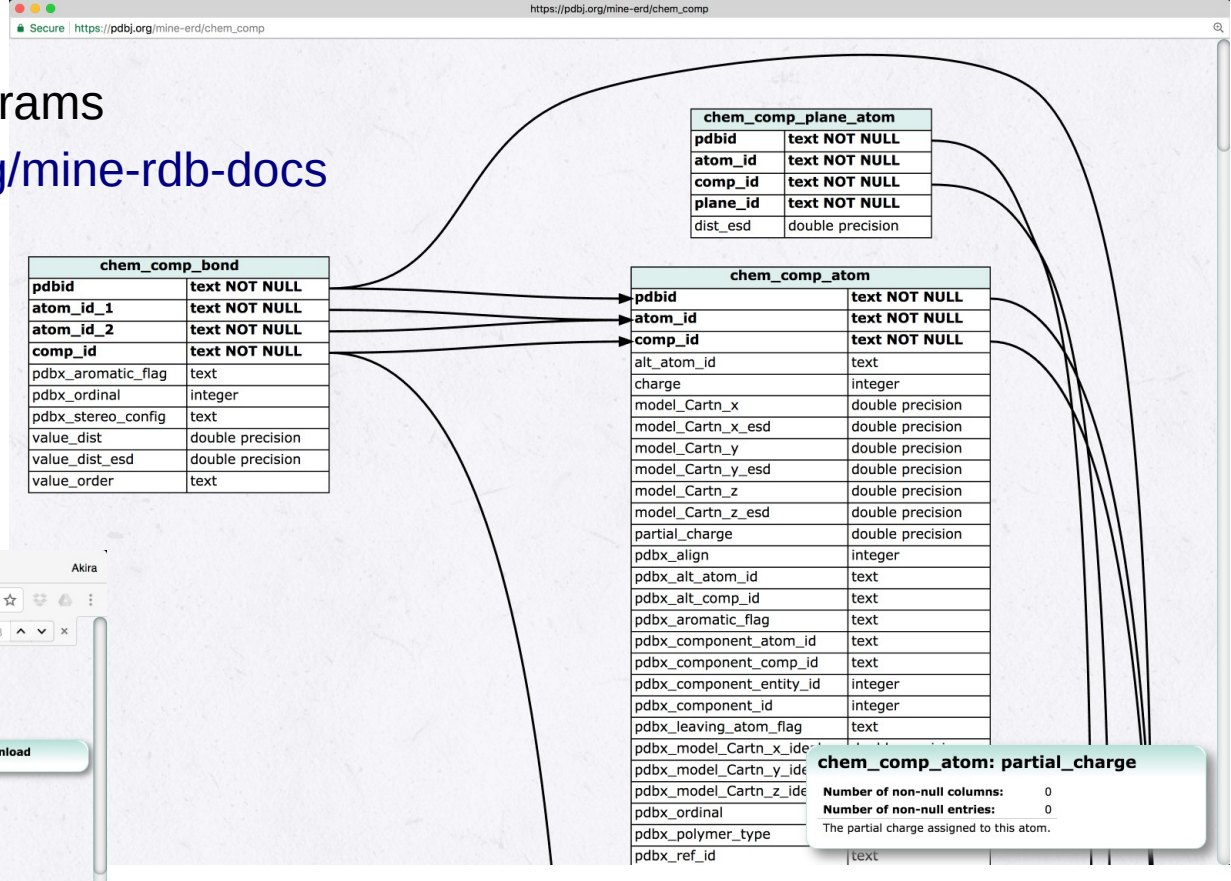
Enter the PDBj Mine2 RDB

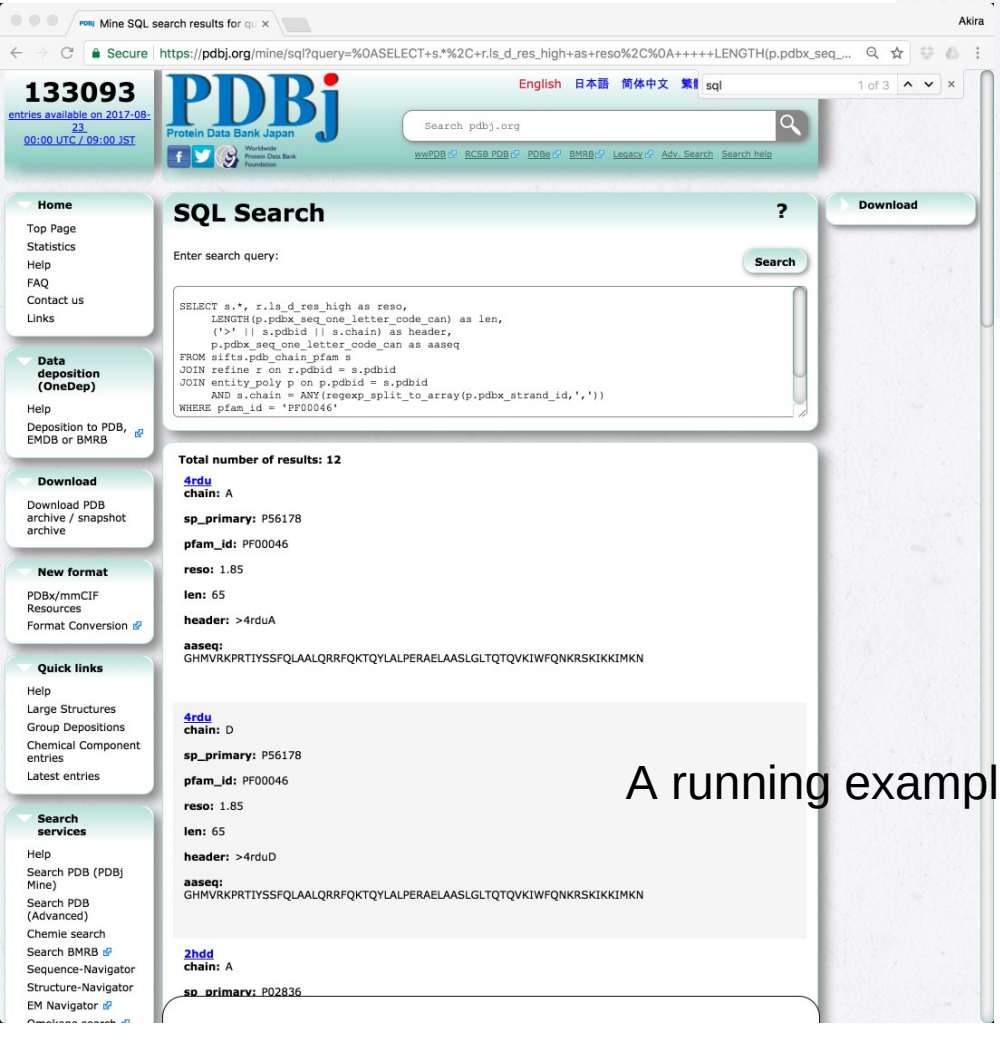
PDBj Mine2 RDB

- Relational database working behind the PDBj.
- Docs: https://pdbj.org/mine-rdb-docs
  - Complete database schema with diagrams.
- Web SQL interface: https://pdbj.org/mine
- REST API: https://pdbj.org/rest/mine2_sql
- SQL dump: ftp://ftp.pdbj.org/mine2/
  - Requires **PostgreSQL** >= 9.3
  - See https://pdbj.org/help/mine2-rdb-local-install
- Many examples: https://pdbj.org/help/mine2-sql

Schema diagrams

https://pdbj.org/mine-rdb-docs





A running example on PDBj web site.

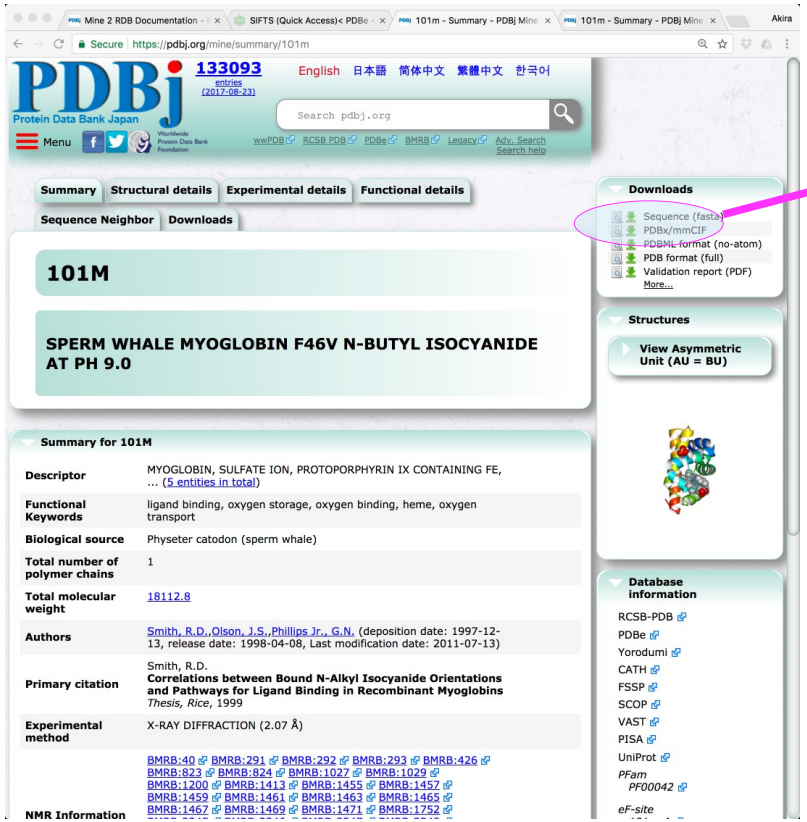
mmCIF-Mine2 RDB mapping

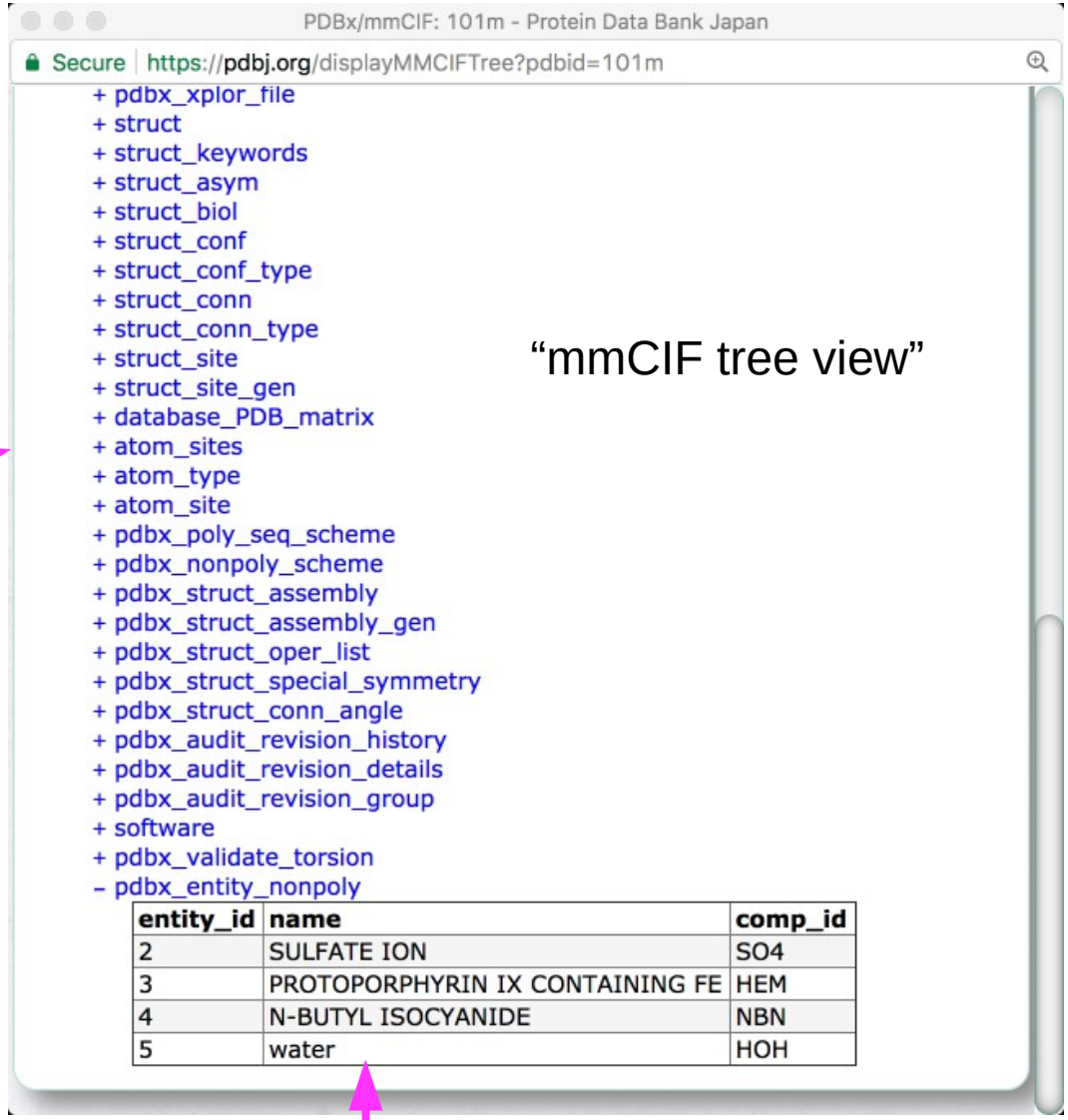
mmCIF	Mine2 RDB
Category	Table
Item	Column
Key data item	Primary key (w/ PDBID)
Parent-child relation	Foreign key

Caution: The mapping is *NOT* exact due to semantic differences.

Ex. PDB entries containing “HEM”

```
SELECT pdbid
FROM pdbx_entity_nonpoly
WHERE comp_id = 'HEM'
```



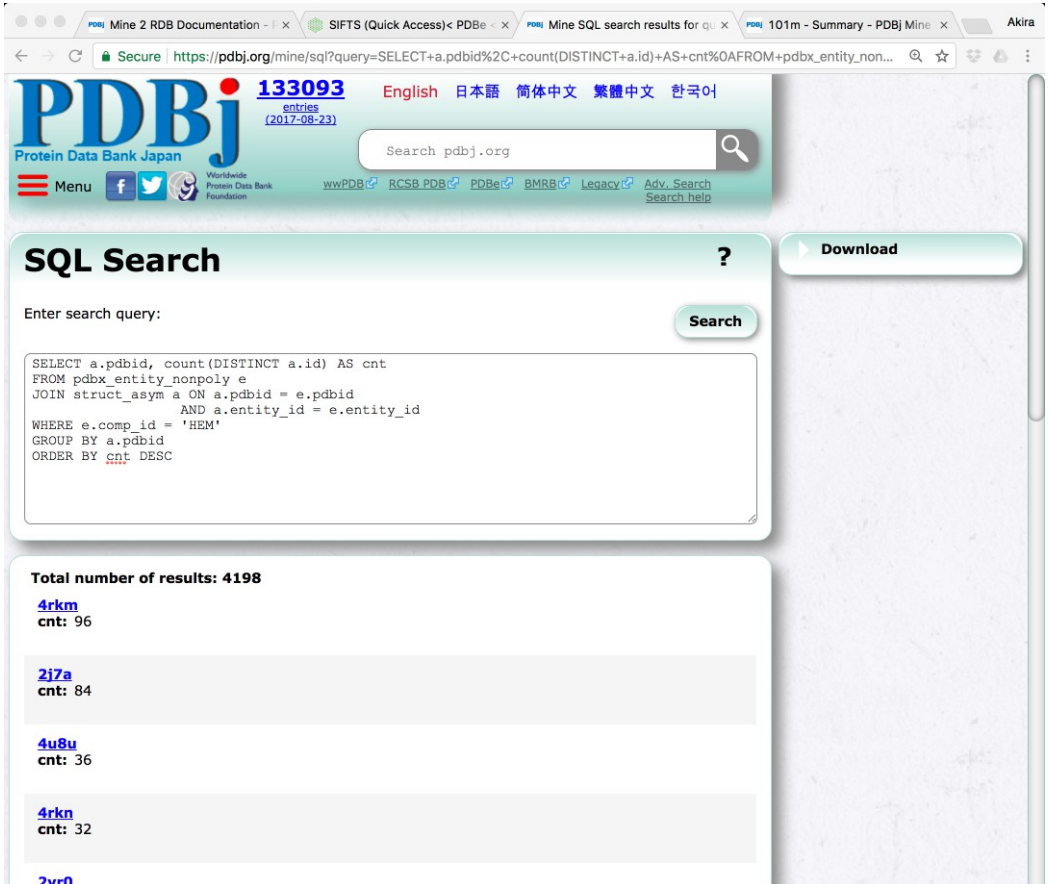


“mmCIF tree view”

“pdbx_entity_nonpoly” category

Ex. PDB entries containing “HEM” sorted by the number of HEM's in asymmetric unit

```
SELECT a.pdbid, count(DISTINCT a.id) AS cnt
FROM pdbx_entity_nonpoly e
JOIN struct_asym a ON a.pdbid = e.pdbid
                    AND a.entity_id = e.entity_id
WHERE e.comp_id = 'HEM'
GROUP BY a.pdbid
ORDER BY cnt DESC
```



New: Group deposition

Group of entries deposited in a collection.

```
SELECT group_id, group_title, COUNT(DISTINCT pdbid) AS cnt
FROM pdbx_deposit_group
GROUP BY group_id, group_title
ORDER BY cnt DESC
```

Total number of results: 16

group_id: G_1002001
group_title: High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits.
cnt: 364

group_id: G_1002023
group_title: PanDDA analysis group deposition
cnt: 257

group_id: G_1002021
group_title: PanDDA analysis group deposition
cnt: 224

group_id: G_1002019
group_title: PanDDA analysis group deposition
cnt: 199

group_id: G_1002025
group_title: PanDDA analysis group deposition
cnt: 114

group_id: G_1002022
group_title: PanDDA analysis group deposition
cnt: 40

group_id: G_1002033
group_title: Ligand binding to FARNESOID-X-RECEPTOR
cnt: 37

Summary for 5OYQ

Group deposition	High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits. (G_1002001)
Descriptor	Endothiapepsin (2 entities in total)
Functional Keywords	fragment screening, method development, aspartic protease, hydrolase
Biological source	Cryphonectria parasitica (Chesnut blight fungus)
Total number of polymer chains	1
Total molecular weight	33813.86
Authors	Schiebel, J.,Heine, A.,Klebe, G. (deposition date: 2016-06-28, release date: 2016-08-03, Last modification date: 2017-01-11)
Primary citation	Schiebel, J.,Krimmer, S.G.,Rower, K.,Knorlein, A.,Wang, X.,Park, A.Y.,Stieler, M.,Ehrmann, F.R.,Fu, K.,Radeva, N.,Krug, M.,Huschmann, F.U.,Glockner, S.,Weiss, M.S.,Mueller, U.,Klebe, G.,Heine, A. **High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits.** *Structure*, 24:1398-1409, 2016 PubMed: 27452405 (PDB entries with the same primary citation) DOI: 10.1016/j.str.2016.06.010 Import into Mendeley
Experimental method	X-RAY DIFFRACTION (1.49 Å)

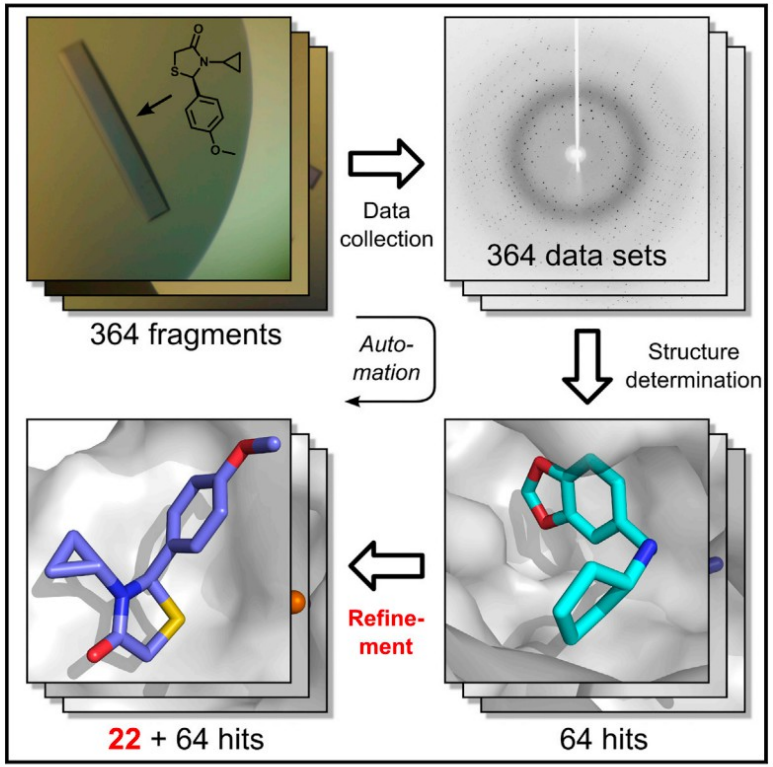
group_id

Resource

Structure **24**:1398–1409, (2016)

High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits

Graphical Abstract



Authors

Johannes Schiebel, Stefan G. Krimmer, Karine Röwer, ..., Uwe Mueller, Gerhard Klebe, Andreas Heine

Correspondence

heinea@mailer.uni-marburg.de

In Brief

Schiebel et al. describe a large collection of diffraction data obtained from protein crystals treated with 364 different fragment molecules. An automated structural refinement of these datasets suggests that more complete refinement enables detection of additional fragments binding in the target's active site.

Highlights

- Reliable fragment hit identification requires phase improvement through refinement
- The modeling of water molecules particularly enhances fragment electron densities
- 364 diffraction datasets are provided for the design and validation of new methods
- An automated procedure was developed to streamline structural refinement

"... exhaustively ref ne all structures before inspecting electron density maps for putatively bound fragments."

BIRD: Biologically Interesting Molecule Reference Dictionary

- See https://www.wwpdb.org/data/bird
- Antibiotics, inhibitors, etc.

In 1KQE:

```
_pdbx_molecule_features.prd_id    PRD_000154
_pdbx_molecule_features.name      'MINI-GRAMICIDIN A DIMER'
_pdbx_molecule_features.type      Polypeptide
_pdbx_molecule_features.class     Antibiotic
_pdbx_molecule_features.details
;THE N-TERMINI OF THE TWO IDENTICAL PEPTIDES, EACH
 A TRUNCATED GRAMICIDIN A WERE LINKED BY A SUCCINIC
 ACID IN A HEAD-TO-HEAD MANNER.
;
#
```

List BIRD entries or their types according to popularity

```
SELECT prd_id, name, COUNT(pdbid)
FROM  pdbx_molecule_features
GROUP BY prd_id,name
ORDER BY COUNT DESC
```

Total number of results: 835

prd_id: PRD_000020
name: D-Phe-Pro-Arg-CH2Cl
count: 51

prd_id: PRD_000238
name: Ac-Asp-Glu-Val-Asp-CMK
count: 46

prd_id: PRD_000142
name: Cyclosporin A
count: 30

prd_id: PRD_000398
name: N-{(2S)-2-[(N-acetyl-L-threonyl-L-isoleucyl)amino]hexyl}-L-norleucyl-L-glutaminyl-N~5~-[amino(iminio)methyl]-L-ornithinamide
count: 29

prd_id: PRD_001243
name: CARFILZOMIB, bound form
count: 28

prd_id: PRD_000454
name: Saquinavir
count: 27

prd_id: PRD_000557
name: Pepstatin
count: 24

```
SELECT type, COUNT(pdbid)
FROM pdbx_molecule_features
GROUP BY type
ORDER BY COUNT DESC
```

Total number of results: 17

type: Peptide-like
count: 1032

type: Oligopeptide
count: 146

type: Cyclic peptide
count: 130

type: Polypeptide
count: 117

type: Glycopeptide
count: 38

type: Cyclic depsipeptide
count: 23

type: Thiopeptide
count: 18

type: Peptaibol
count: 15

type: Non-polymer
count: 8

type: Cyclic lipopeptide
count: 3

type: Lipopeptide
count: 3

Integration with SIFTS

- “Structure Integration with Function, Taxonomy and Sequence” developed by PDBe & UniProt.
  - https://www.ebi.ac.uk/pdbe/docs/sifts/
- Integrates *UniProt*, *NCBI Taxonomy*, *Gene Ontology*, *Pfam*, *EC code*, *PubMed*, *SCOP*, *CATH* with PDB.
- The “Quick access” data of SIFTS are integrated into the PDBj Mine RDB.
  - c.f., https://www.ebi.ac.uk/pdbe/docs/sifts/quick.html

Assign UniProt accession to each entry

```
SELECT e.pdbid, e.entity_id, s.chain, s.sp_primary
FROM entity_poly e
JOIN sifts.pdb_chain_uniprot s
  ON s.pdbid = e.pdbid
   AND s.chain = ANY(STRING_TO_ARRAY(e.pdbx_strand_id,',' ))
```

* Sometimes Postgres-specific functions are handy.

* “entity_poly.pdbx_strand_id” provides a comma-separated list of auth_asym_id's (a.k.a. chain ID's).

In 1A00:

```
loop_
_entity_poly.entity_id
_entity_poly.type
_entity_poly.nstd_linkage
_entity_poly.nstd_monomer
_entity_poly.pdbx_seq_one_letter_code
_entity_poly.pdbx_seq_one_letter_code_can
_entity_poly.pdbx_strand_id
1 'polypeptide(L)' no no
;VLSPADKTNVKAAWGKVGAHAGEYGAEALE.....
;
;VLSPADKTNVKAAWGKVGAHAGEYGAEALERM.....
;
A,C
2 'polypeptide(L)' no no
;MHLTPEEKSAVTALWGKVNVDEVGGEALGRLLV.....
;
;MHLTPEEKSAVTALWGKVNVDEVGGEALGRLLV.....
B,D
#
```

Find polymer entities with GO:0006220 (pyrimidine nucleotide metabolic process) in human

```
SELECT DISTINCT e.pdbid, e.entity_id, s.sp_primary
FROM entity_poly e
JOIN sifts.pdb_chain_go s
  ON s.pdbid = e.pdbid
  AND s.chain = ANY(STRING_TO_ARRAY(e.pdbx_strand_id,','))
JOIN sifts.pdb_chain_taxonomy t
  ON t.pdbid = e.pdbid AND t.chain = s.chain
WHERE s.go_id = 'GO:0006220'
  AND t.scientific_name = 'Homo sapiens'
```

Find polymer entities assigned to multiple species (i.e., chimeric proteins)

```
SELECT e.pdbid,e.entity_id, COUNT(DISTINCT t.tax_id) AS n
     , ARRAY_AGG(DISTINCT t.tax_id) AS tax_ids
FROM entity_poly e
JOIN sifts.pdb_chain_taxonomy t
  ON t.pdbid = e.pdbid
  AND t.chain = ANY(STRING_TO_ARRAY(e.pdbx_strand_id,','))
GROUP BY e.pdbid,e.entity_id
HAVING COUNT(DISTINCT t.tax_id) > 1
ORDER BY n DESC
```

Retrieve PDB chain sequences matching to the Pfam accession "PF00046" (Homeobox) and having resolution better than 2.0 Angstrom and sequence length greater than or equal to 58 (residues)

```
SELECT s.*, r.ls_d_res_high as reso,
     LENGTH(p.pdbx_seq_one_letter_code_can) as len,
     ('>' || s.pdbid || s.chain) as header,
     p.pdbx_seq_one_letter_code_can as aaseq
FROM sifts.pdb_chain_pfam s
JOIN refine r on r.pdbid = s.pdbid
JOIN entity_poly p on p.pdbid = s.pdbid
     AND s.chain = ANY(regexp_split_to_array(p.pdbx_strand_id,','))
WHERE pfam_id = 'PF00046'
AND r.ls_d_res_high < 2.0
AND LENGTH(p.pdbx_seq_one_letter_code_can) >= 58
ORDER BY reso, len,s.chain
```

* “entity_poly.pdbx_seq_one_letter_code_can” contains AA seq in one-letter codes.

Mine SQL search resul | SIFTS (Quick Access) | Mine SQL search resul | Mine SQL search resul | 1a00 - Summary - PDB | Akira

Secure | https://pdbj.org/mine/sql?query=%0ASELECT+s.*%2C+r.ls_d_res_high+as+reso%2C%0A+++++LENGTH(p.pdbx_se...

PDBj Protein Data Bank Japan — Worldwide Protein Data Bank Foundation

133093 entries (2017-08-23)

English 日本語 简体中文 繁體中文 한국어

Search pdbj.org

Menu | wwPDB | RCSB PDB | PDBe | BMRB | Legacy | Adv. Search | Search help

SQL Search

?

Download

Enter search query:

Search

```
SELECT s.*, r.ls_d_res_high as reso,
     LENGTH(p.pdbx_seq_one_letter_code_can) as len,
     ('>' || s.pdbid || s.chain) as header,
     p.pdbx_seq_one_letter_code_can as aaseq
FROM sifts.pdb_chain_pfam s
JOIN refine r on r.pdbid = s.pdbid
JOIN entity_poly p on p.pdbid = s.pdbid
     AND s.chain = ANY(regexp_split_to_array(p.pdbx_strand_id,','))
WHERE pfam_id = 'PF00046'
AND r.ls_d_res_high < 2.0
AND LENGTH(p.pdbx_seq_one_letter_code_can) >= 58
```

Total number of results: 12

4rdu
chain: A

sp_primary: P56178

pfam_id: PF00046

reso: 1.85

len: 65

header: >4rduA

aaseq:
GHMVRKPRTIYSSFQLAALQRRFQKTQYLALPERAELAASLGLTQTQVKIWFQNKRSKIKKIMKN

4rdu
chain: D

sp_primary: P56178

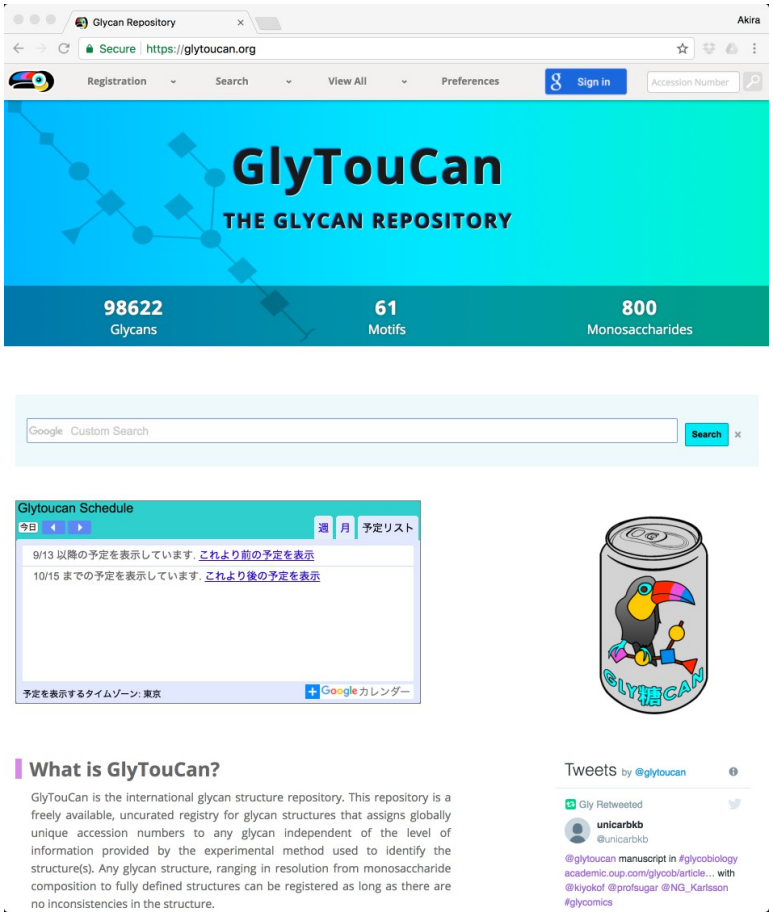
pfam_id: PF00046

reso: 1.85

Integration with GlyTouCan

Listing PDB ID's that contain GlyTouCan entries.

```
SELECT cc.pdbid, g.acc
FROM glytoucan.chem_comp g
JOIN chem_comp cc ON cc.id = g.chem_comp_id
```



https://glytoucan.org

"GlyTouCan is the international glycan structure repository."

Coming soon!

More integration with:

- Chemical Component Dictionary (cc)
  - PDB's (3-letter) chemical components
  - Includes InChi keys, SMILES, etc.
- Chemical Component Model Data (ccmodel)
  - Xref to *Cambridge Structure Database* (CSD)
- BIRD
  - “Biologically Interesting Molecule Reference Dictionary”
  - Peptide-like antibiotic and inhibitor molecules.

If you want to do complicated queries,
we may be able to help!

Feel free to ask any questions at:
https://pdbj.org/contact?tab=PDBjmaster