PDB の「新」フォーマット

金城玲

日本蛋白質構造データバンク (PDBj)
大阪大学蛋白質研究所

概要

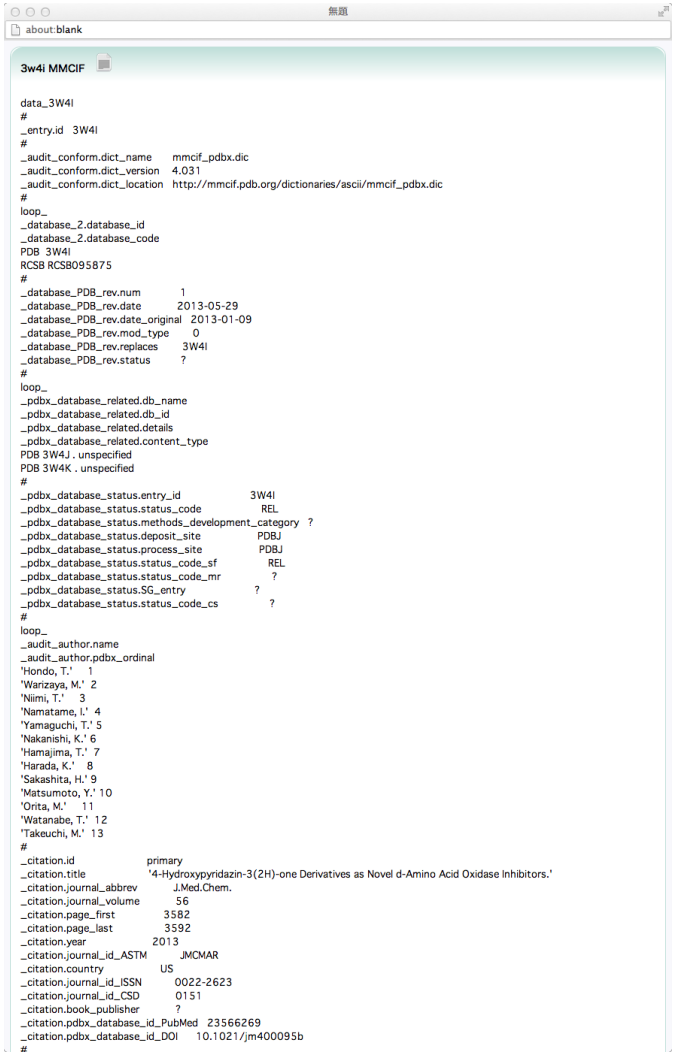
- PDB の「新」フォーマットとは？
  - PDBx/mmCIF
- なぜ「新」フォーマットに移行するのか？
  - PDB format の限界
- 対応済みソフトウェア
- 自作のソフトはどうする？

「新」フォーマットとは？

PDBx/mmCIF のことです！

- PDBx (PDB exchange dictionary) に基づく。
- STAR(Self-defining Text Archive and Retrieval) 文法で記述されている。
- CIF(Crystallographic Information Format) を PDB 用に拡張したもの： macromolecular CIF → mmCIF

PDBx/mmCIF を見てみる



もう少し良く見てみる

3w4i MMCIF

```
data_3W4I
#
_entry.id   3W4I
#
_audit_conform.dict_name      mmcif_pdbx.dic
_audit_conform.dict_version   4.031
_audit_conform.dict_location  http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB  3W4I
RCSB RCSB095875
#
_database_PDB_rev.num            1
_database_PDB_rev.date           2013-05-29
_database_PDB_rev.date_original  2013-01-09
_database_PDB_rev.mod_type       0
_database_PDB_rev.replaces       3W4I
_database_PDB_rev.status         ?
#
loop_
_pdbx_database_related.db_name
_pdbx_database_related.db_id
_pdbx_database_related.details
_pdbx_database_related.content_type
PDB 3W4J . unspecified
PDB 3W4K . unspecified
#
```

....

....

datablock

entry ID

繰返し項目 (loop)

```
_citation.id                    primary
_citation.title                 '4-Hydroxypyridazin-3(2H)-one Derivatives as Novel d-Amino Acid Oxidase Inhibitors.'
_citation.journal_abbrev        J.Med.Chem.
_citation.journal_volume        56
_citation.page_first            3582
_citation.page_last             3592
_citation.year                  2013
_citation.journal_id_ASTM       JMCMAR
_citation.country               US
_citation.journal_id_ISSN       0022-2623
_citation.journal_id_CSD        0151
_citation.book_publisher        ?
_citation.pdbx_database_id_PubMed   23566269
_citation.pdbx_database_id_DOI      10.1021/jm400095b
#
loop_
_citation_author.citation_id
_citation_author.name
_citation_author.ordinal
primary 'Hondo, T.'     1
primary 'Warizaya, M.'  2
primary 'Niimi, T.'     3
primary 'Namatame, I.'  4
primary 'Yamaguchi, T.' 5
primary 'Nakanishi, K.' 6
primary 'Hamajima, T.'  7
primary 'Harada, K.'    8
primary 'Sakashita, H.' 9
primary 'Matsumoto, Y.' 10
primary 'Orita, M.'     11
primary 'Takeuchi, M.'  12
#
_cell.entry_id          3W4I
_cell.length_a          148.696
_cell.length_b          181.529
_cell.length_c          50.612
_cell.angle_alpha       90.00
_cell.angle_beta        90.00
_cell.angle_gamma       90.00
_cell.Z_PDB             16
_cell.pdbx_unique_axis  ?
_cell.length_a_esd      ?
_cell.length_b_esd      ?
_cell.length_c_esd      ?
_cell.angle_alpha_esd   ?
_cell.angle_beta_esd    ?
_cell.angle_gamma_esd   ?
#
```

PDBx/mmCIF の基本

- データはいろいろなカテゴリに分類されている
  - *_category.item*
  - 例： _entry.id → “_entry” はカテゴリ名、 “ id” はその項目 (item)
  - 「 _entry.id　3W4I 」は entry カテゴリの id 項目の値が 3W4I である、という意味。
- データの記述の仕方は２通り
  - key-value: 一つのカテゴリに一つの値しかない場合
  - loop: 一つのカテゴリに複数の値がある場合

key-value の例

```
_cell.entry_id              3W4I
_cell.length_a              148.696
_cell.length_b              181.529
_cell.length_c              50.612
_cell.angle_alpha           90.00
_cell.angle_beta            90.00
_cell.angle_gamma           90.00
_cell.Z_PDB                 16
_cell.pdbx_unique_axis      ?
_cell.length_a_esd          ?
_cell.length_b_esd          ?
_cell.length_c_esd          ?
_cell.angle_alpha_esd       ?
_cell.angle_beta_esd        ?
_cell.angle_gamma_esd       ?
#
```

最後の「#」はそのカテゴリの記述の終わりを表す convention

loop の例

```
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 'D-amino-acid oxidase'        39521.289 4 ? ? ? 1.4.3.3
2 non-polymer syn 'FLAVIN-ADENINE DINUCLEOTIDE' 785.557   4 ? ? ? ?
3 non-polymer syn PYRIDINE-2,3-DIOL             111.100   4 ? ? ? ?
#
```

ループの開始

項目のリスト

実際のデータ

- 各項目は空白で区切られる
- 項目リストと同じ順番で並ぶ
- 空白を含むデータは引用府「'」で囲む

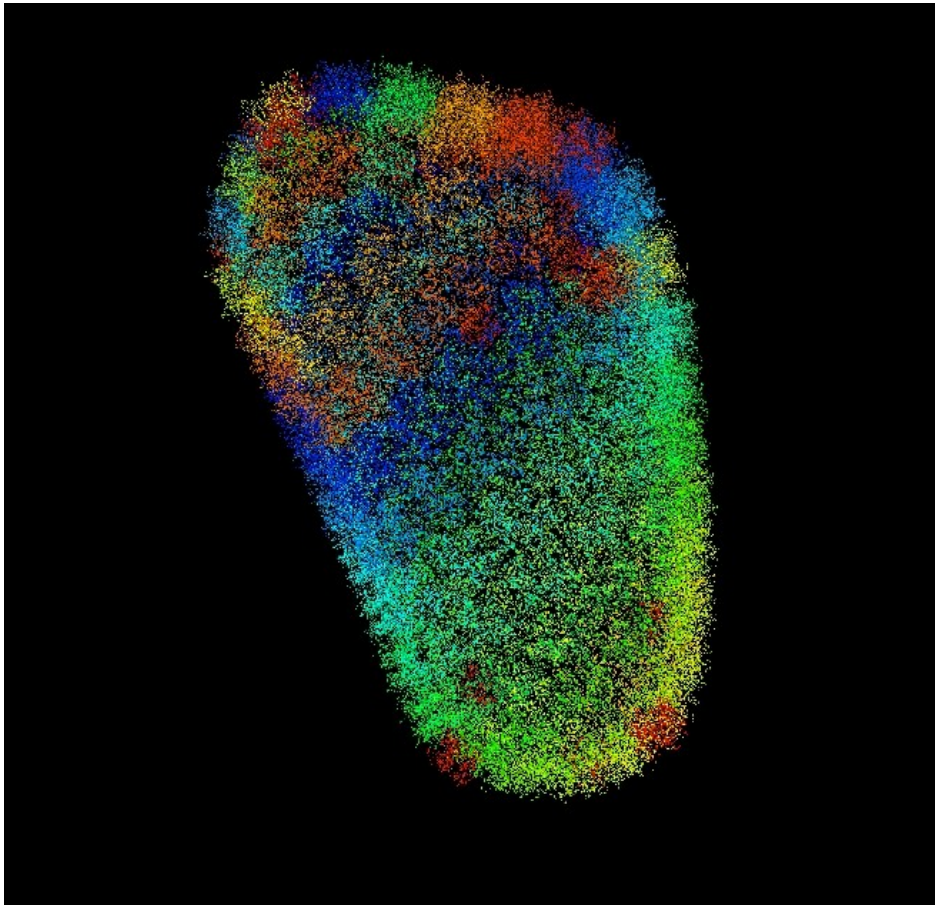
最後の「#」はそのループの終わりを表す convention

なぜ PDBx/mmCIF を使うか？

PDB フォーマットは もう限界です！

- 固定コラム数
  - 最大 99,999 原子まで
  - 最大 36 chain まで
    - 現実には「反則ワザ」でもう少し増やしていますが……
  - 座標は最大４桁まで（負号がある場合は３桁まで）
- アノテーションの不完全さ
  - 複雑怪奇な REMARK 行の自動処理は、「例外」処理がルーチン化している
  - 残基番号の一貫性がない
  - 外部 DB との連携が難しい

巨大構造の例



- HIV-1 capsid (3J3Q 他)
  - 1,356 鎖
  - 2,440,800 原子
  - 25 PDB エントリ
    - 1VU5,1VU6,...
  - 「3J3Q」にまとめられている
    - mmCIF, PDBML のみ

その他の巨大構造について

http://mmcif.pdb.org/large-pdbx-examples/

ftp://ftp.pdbj.org/pub/pdb/data/large_structures/mmCIF/
ftp://ftp.pdbj.org/pub/pdb/data/large_structures/XML/

「ATOM 行」の例（整合性）

```
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     . MET A 1 1   ? 38.130  3.618   -16.026 1.00 62.74 ? ? ? ? ? ? 1   MET A N    1
ATOM   2     C CA    . MET A 1 1   ? 38.828  4.206   -14.836 1.00 62.58 ? ? ? ? ? ? 1   MET A CA   1
ATOM   3     C C     . MET A 1 1   ? 37.812  4.785   -13.869 1.00 61.80 ? ? ? ? ? ? 1   MET A C    1
ATOM   4     O O     . MET A 1 1   ? 37.403  5.938   -13.998 1.00 61.77 ? ? ? ? ? ? 1   MET A O    1
ATOM   5     C CB    . MET A 1 1   ? 39.811  5.286   -15.292 1.00 63.05 ? ? ? ? ? ? 1   MET A CB   1
ATOM   6     C CG    . MET A 1 1   ? 40.786  5.768   -14.235 1.00 65.31 ? ? ? ? ? ? 1   MET A CG   1
ATOM   7     S SD    . MET A 1 1   ? 41.764  4.451   -13.470 1.00 70.00 ? ? ? ? ? ? 1   MET A SD   1
ATOM   8     C CE    . MET A 1 1   ? 40.683  4.024   -12.124 1.00 70.52 ? ? ? ? ? ? 1   MET A CE   1
ATOM   9     N N     . ARG A 1 2   ? 37.368  3.972   -12.920 1.00 60.73 ? ? ? ? ? ? 2   ARG A N    1
ATOM   10    C CA    . ARG A 1 2   ? 36.391  4.441   -11.938 1.00 59.98 ? ? ? ? ? ? 2   ARG A CA   1
```

PDB で正規化されたデータ

対応

登録者が指定したデータ

PDBx/mmCIF 対応ソフトウェア

- 分子グラフィクス
  - CHEMERA
  - Jmol
  - OpenRasMol
- 構造決定
  - CCP4
  - Phenix

自作ソフトはどうしたら良い？

- ライブラリを使う
  - C/C++, Java, Perl, Python などが既に使える。
- 自分でパーサを書く
  - STAR 形式の BNF 文法を勉強する。
- ものぐさな人は……
  - 今まで PDB フォーマットを扱っていた要領で、「慣習」を利用することもできる。
  - ATOM 行だけなら比較的簡単です。
- または、 PDBML(XML) を使う。

In 2014...

- “Large structure” without PDB files.
- 対応する split エントリはリリースされない。
- 現在の split エントリは削除され、あらたに large structure としてまとめられて再リリースされる。

http://mmcif.pdb.org/large-pdbx-examples/

のデータでソフトのテストをして下さい。

参考文献

- S. R. Hall, *J. Chem. Inf. Comput. Sci.* (1991) **31**, 326-333 (STAR 形式の概要)
- S. R. Hall, *J. Chem. Inf. Comput. Sci.* (1994) **34**, 505-508 (STAR 形式の仕様)
- http://mmcif.pdb.org/ (mmCIF 関連情報の拠点)
- *International Tables for Crystallography G*
  - Ch. 3.6 ... mmCIF の基本概念など
- PDBj 講習会の資料など
  - 「 mmCIF と PDBML 」