PDB データの読み解き方
mmCIF と PDBML

金城玲
大阪大学蛋白質研究所

この講義の内容

- PDB の「新しい」フォーマットの必要性
- PDB の「新しい」フォーマットである mmCIF
- その XML への「直訳」である PDBML
- mmCIF と PDBML の基礎である PDBx 辞書
- これらの具体例

PDB の「新」フォーマット

- 「旧」フォーマットは、いわゆる PDB ファイルのフォーマットのこと。
- 「新」フォーマットは、 mmCIF ファイルのフォーマットのこと。（実は新しくない）
- 2014 年から本格的に移行します。

従来の PDB フォーマットの例

```
HEADER    NEUROPEPTIDE                                  27-JUL-12   2LWC
TITLE     MET-ENKEPHALIN IN DPMC SUV
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MET-ENKEPHALIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 SYNTHETIC: YES;
SOURCE   3 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   4 ORGANISM_COMMON: HUMAN;
SOURCE   5 ORGANISM_TAXID: 9606
KEYWDS    SUV DMPC, NEUROPEPTIDE
EXPDTA    SOLUTION NMR
NUMMDL    20
…
MODEL        1
ATOM      1  N   TYR A   1      -4.388  -4.949   1.122  1.00  1.00           N
ATOM      2  CA  TYR A   1      -4.868  -4.392  -0.176  1.00  1.00           C
ATOM      3  C   TYR A   1      -3.679  -3.966  -1.041  1.00  1.00           C
ATOM      4  O   TYR A   1      -2.544  -3.995  -0.611  1.00  1.00           O
ATOM      5  CB  TYR A   1      -5.716  -3.178   0.205  1.00  1.00           C
ATOM      6  CG  TYR A   1      -6.423  -2.655  -1.023  1.00  1.00           C
ATOM      7  CD1 TYR A   1      -7.091  -3.540  -1.877  1.00  1.00           C
ATOM      8  CD2 TYR A   1      -6.409  -1.283  -1.307  1.00  1.00           C
ATOM      9  CE1 TYR A   1      -7.746  -3.054  -3.014  1.00  1.00           C
ATOM     10  CE2 TYR A   1      -7.064  -0.798  -2.445  1.00  1.00           C
…
```

mmCIF の例

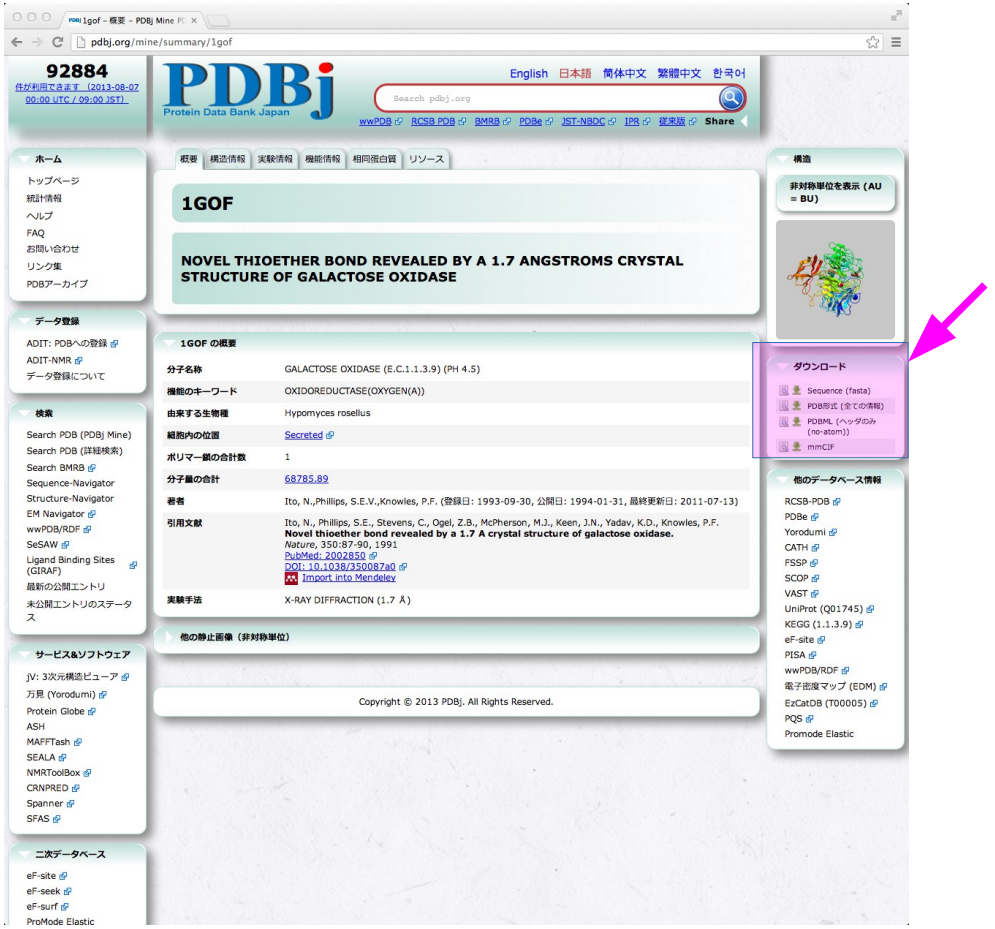
```
data_12AS
#
_entry.id   12AS
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    4.007
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
_database_2.database_id     PDB
_database_2.database_code   12AS
#
loop_
_database_PDB_rev.num
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.status
_database_PDB_rev.replaces
_database_PDB_rev.mod_type
1 1998-12-30 1997-12-02 ? 12AS 0
2 1999-02-16 ?          ? 12AS 3
3 2009-02-24 ?          ? 12AS 1
…
```

mmCIF の例（続き）

```
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
…
```

実際に見てみる

PDB ID



「PDB 形式」と「mmCIF」を見る

PDB形式 (全ての情報): 1gof – 日本蛋白質構造データバンク

pdbj.org/displayPDBfile?mD=1&pdbid=1gof

PDB形式 (全ての情報): 1gof

```
HEADER    OXIDOREDUCTASE(OXYGEN(A))          30-SEP-93   1GOF
TITLE     NOVEL THIOETHER BOND REVEALED BY A 1.7 ANGSTROMS CRYSTAL STRUCTURE OF
TITLE    2 GALACTOSE OXIDASE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GALACTOSE OXIDASE;
COMPND   3 CHAIN: A;
COMPND   4 EC: 1.1.3.9;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HYPOMYCES ROSELLUS;
SOURCE   3 ORGANISM_TAXID: 5132
KEYWDS    OXIDOREDUCTASE(OXYGEN(A))
EXPDTA    X-RAY DIFFRACTION
AUTHOR    N.ITO,S.E.V.PHILLIPS,P.F.KNOWLES
REVDAT   4  13-JUL-11 1GOF   1      VERSN
REVDAT   3  24-FEB-09 1GOF   1      VERSN
REVDAT   2  01-APR-03 1GOF   1      JRNL
REVDAT   1  31-JAN-94 1GOF   0
JRNL      AUTH  N.ITO,S.E.PHILLIPS,C.STEVENS,Z.B.OGEL,M.J.MCPHERSON,
JRNL      AUTH 2 J.N.KEEN,K.D.YADAV,P.F.KNOWLES
JRNL      TITL  NOVEL THIOETHER BOND REVEALED BY A 1.7 A CRYSTAL STRUCTURE
JRNL      TITL 2 OF GALACTOSE OXIDASE.
JRNL      REF   NATURE                V. 350   87 1991
JRNL      REFN              ISSN 0028-0836
JRNL      PMID  2002850
JRNL      DOI   10.1038/350087A0
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH  N.ITO,S.E.V.PHILLIPS,K.K.S.YADAV,P.F.KNOWLES
REMARK   1  TITL  THE CRYSTAL STRUCTURE OF A FREE RADICAL ENZYME, GALACTOSE
REMARK   1  TITL 2 OXIDASE
REMARK   1  REF   TO BE PUBLISHED
REMARK   1  REFN
REMARK   1 REFERENCE 2
REMARK   1  AUTH  M.J.MCPHERSON,Z.B.OGEL,C.STEVENS,K.D.S.YADAV,J.M.KEEN,
REMARK   1  AUTH 2 P.F.KNOWLES
REMARK   1  TITL  GALACTOSE OXIDASE OF DACTYLIUM DENDROIDES: GENE CLONING AND
REMARK   1  TITL 2 SEQUENCE ANALYSIS
REMARK   1  REF   J.BIOL.CHEM.              V. 267  8146 1992
REMARK   1  REFN              ISSN 0021-9258
REMARK   2
REMARK   2 RESOLUTION.    1.70 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3  PROGRAM     : PROLSQ
REMARK   3  AUTHORS     : KONNERT,HENDRICKSON
REMARK   3
REMARK   3 DATA USED IN REFINEMENT.
```

mmCIF: 1gof – 日本蛋白質構造データバンク

pdbj.org/displayMMCIF?mD=1&pdbid=1gof

mmCIF: 1gof

```
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
#
loop_
_database_PDB_rev.num
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.status
_database_PDB_rev.replaces
_database_PDB_rev.mod_type
1 1994-01-31 1993-09-30 ? 1GOF 0
2 2003-04-01 ?       ? 1GOF 1
3 2009-02-24 ?       ? 1GOF 1
4 2011-07-13 ?       ? 1GOF 1
#
loop_
_database_PDB_rev_record.rev_num
_database_PDB_rev_record.type
_database_PDB_rev_record.details
2 JRNL  ?
3 VERSN ?
4 VERSN ?
#
_pdbx_database_status.status_code     REL
_pdbx_database_status.entry_id       1GOF
_pdbx_database_status.deposit_site    ?
_pdbx_database_status.process_site    ?
_pdbx_database_status.status_code_sf  REL
_pdbx_database_status.status_code_mr  ?
_pdbx_database_status.SG_entry       ?
#
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Ito, N.'        1
'Phillips, S.E.V.' 2
'Knowles, P.F.'   3
#
loop_
_citation.id
```

mmCIF をもう少し見てみる

```
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
#
loop_
_database_PDB_rev.num
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.status
_database_PDB_rev.replaces
_database_PDB_rev.mod_type
1 1994-01-31 1993-09-30 ? 1GOF 0
2 2003-04-01 ?       ? 1GOF 1
3 2009-02-24 ?       ? 1GOF 1
4 2011-07-13 ?       ? 1GOF 1
#
loop_
_database_PDB_rev_record.rev_num
_database_PDB_rev_record.type
_database_PDB_rev_record.details
2 JRNL  ?
3 VERSN ?
4 VERSN ?
```

- data_1GOF ← datablock
- _entry.id 1GOF ← Entry ID
- loop_ ← 繰り返し項目 (loop)

```
loop_
_citation.id
_citation.title
_citation.journal_abbrev
_citation.journal_volume
_citation.page_first
_citation.page_last
_citation.year
_citation.journal_id_ASTM
_citation.country
_citation.journal_id_ISSN
_citation.journal_id_CSD
_citation.book_publisher
_citation.pdbx_database_id_PubMed
_citation.pdbx_database_id_DOI
primary 'Novel thioether bond revealed by a 1.7 A crystal structure of galactose oxidase.' Nature        350 87   90 1991
NATUAS UK 0028-0836 0006 ? 2002850 10.1038/350087a0
1     'The Crystal Structure of a Free Radical Enzyme, Galactose Oxidase'          'To be Published' ?  ?   ? ?   ? ?
?      0353 ? ?     ?
2     'Galactose Oxidase of Dactylium Dendroides: Gene Cloning and Sequence Analysis'   J.Biol.Chem.     267 8146 ?  1992
JBCHA3 US 0021-9258 0071 ? ?     ?
#
loop_
_citation_author.citation_id
_citation_author.name
_citation_author.ordinal
primary 'Ito, N.'       1
primary 'Phillips, S.E.'  2
primary 'Stevens, C.'     3
primary 'Ogel, Z.B.'     4
primary 'McPherson, M.J.'  5
primary 'Keen, J.N.'     6
primary 'Yadav, K.D.'     7
primary 'Knowles, P.F.'   8
1     'Ito, N.'       9
1     'Phillips, S.E.V.' 10
1     'Yadav, K.K.S.'   11
1     'Knowles, P.F.'   12
2     'McPherson, M.J.'  13
2     'Ogel, Z.B.'     14
2     'Stevens, C.'     15
2     'Yadav, K.D.S.'   16
2     'Keen, J.M.'     17
2     'Knowles, P.F.'   18
#
_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  ?
```

mmCIF の基本

- データは色々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 1GOF 」は entry カテゴリの id 項目の値が「 1GOF 」である、という意味。
- データの記述法は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                  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          ?
#
```

最後の「#」はそのカテゴリの記述の終わりを示す慣習 (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 '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 ? ? ? ?
#
```

ループの開始

項目のリスト（「1行1項目」は慣習）

- 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「#」はループの終わりを示す慣習 (convention)

PDBx: PDB exchange dictionary

- mmCIF のカテゴリや項目は PDBx で定められている。
- PDBx では、項目のデータ型や項目間の依存関係も記述されている。
- http://mmcif.pdb.org/dictionaries/mmcif_pdbx_v40.dic/Index/

  を参照のこと。

主なカテゴリ (グループ)

- _entity　　研究対象の分子情報
  - entity, entity_poly, pdbx_entity_nonpoly, ...
- _atom　　各原子の情報（座標など）
  - atom_site
- _struct　　構造の特色（分子全体、２次構造など）
  - struct, struct_conf, struct_sheet, struct_conn, pdbx_struct_assembly, ...
- _chem_comp　化合物データ
  - chem_comp
- _citation　　文献情報
  - citation, citation_author, ...

カテゴリ間の関係

```
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 ?
#
```

「子 (child) 」

「親子関係」も PDBx で定義されている

「親 (parent) 」

```
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” と “ 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_...”: wwPDB が内部的に付与したラベル

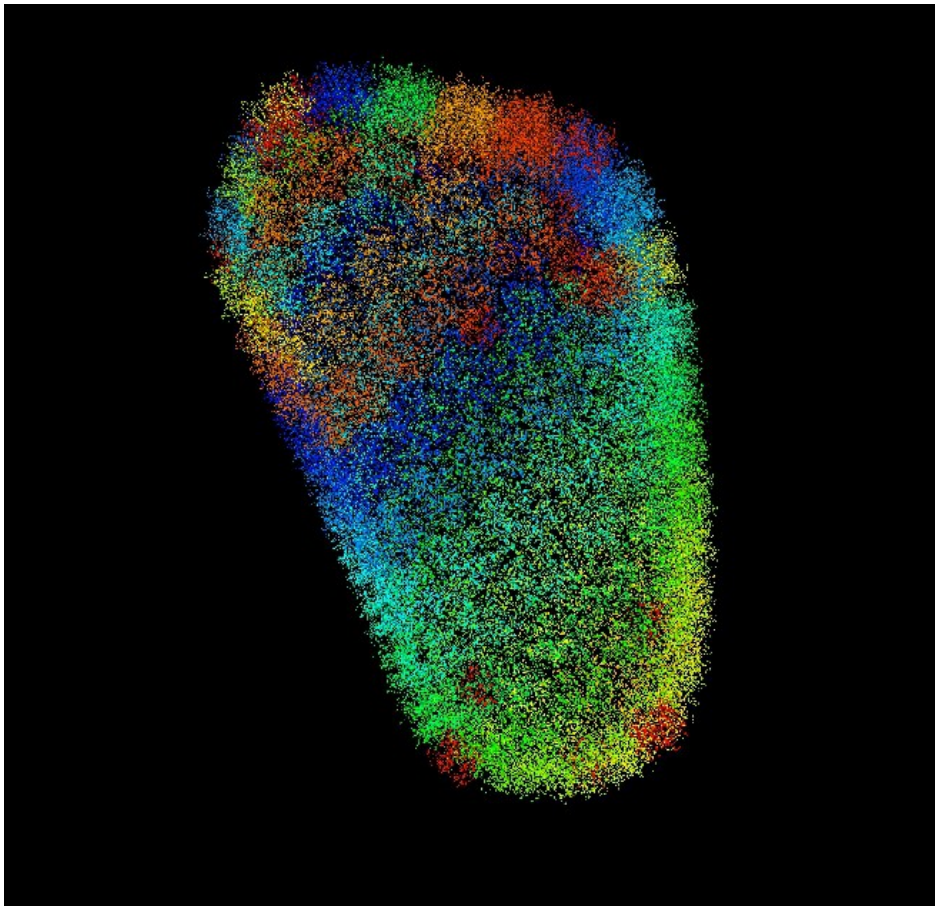
“auth_...”: 登録者が任意に付与したラベル

なぜ PDBx/mmCIF へ完全移行するのか？

PDB フォーマットの限界

- 固定コラム長：大きさの限界
  - 最大 99,999 原子まで。
  - 最大 36 chain まで。（但し、反則ワザあり）
  - 座標は最大 4 桁（負号がある場合は 3 桁）まで。
- アノテーションの不完全さ
  - 複雑奇怪な REMARK 行の自動処理では「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。
  - 外部データベースとの連携が難しい。

巨大構造の例



- HIV-1 capsid (3J3Q)
  - 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/

PDB ファイルのアノテーション

```
JRNL        AUTH   N.ITO,S.E.PHILLIPS,C.STEVENS,Z.B.OGEL,M.J.MCPHERSON,
JRNL        AUTH 2 J.N.KEEN,K.D.YADAV,P.F.KNOWLES
JRNL        TITL   NOVEL THIOETHER BOND REVEALED BY A 1.7 A CRYSTAL STRUCTURE
JRNL        TITL 2 OF GALACTOSE OXIDASE.
JRNL        REF    NATURE                        V. 350    87 1991
JRNL        REFN                   ISSN 0028-0836
JRNL        PMID   2002850
JRNL        DOI    10.1038/350087A0
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   N.ITO,S.E.V.PHILLIPS,K.K.S.YADAV,P.F.KNOWLES
REMARK   1  TITL   THE CRYSTAL STRUCTURE OF A FREE RADICAL ENZYME, GALACTOSE
REMARK   1  TITL 2 OXIDASE
REMARK   1  REF    TO BE PUBLISHED
REMARK   1  REFN
REMARK   1 REFERENCE 2
REMARK   1  AUTH   M.J.MCPHERSON,Z.B.OGEL,C.STEVENS,K.D.S.YADAV,J.M.KEEN,
REMARK   1  AUTH 2 P.F.KNOWLES
REMARK   1  TITL   GALACTOSE OXIDASE OF DACTYLIUM DENDROIDES: GENE CLONING AND
REMARK   1  TITL 2 SEQUENCE ANALYSIS
REMARK   1  REF    J.BIOL.CHEM.                  V. 267  8146 1992
REMARK   1  REFN                   ISSN 0021-9258
REMARK   2
REMARK   2 RESOLUTION.    1.70 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : PROLSQ
REMARK   3   AUTHORS     : KONNERT,HENDRICKSON
REMARK   3
```

注釈の種類により異なる文法

→機械的処理が大変です！

mmCIF の単純な解決法

- 巨大構造
  → 任意長のコラム数、空白区切り
- アノテーション
  → 記述形式は key-value か loop のみ！

PDBML: XML 形式の mmCIF

- PDBx 辞書→ PDBML Schema
- mmCIF→PDBML

データファイル	形式	定義
mmCIF	STAR	PDBx mmCIF dictionary
PDBML	XML	PDBx PDBML Schema

PDBML を使う理由

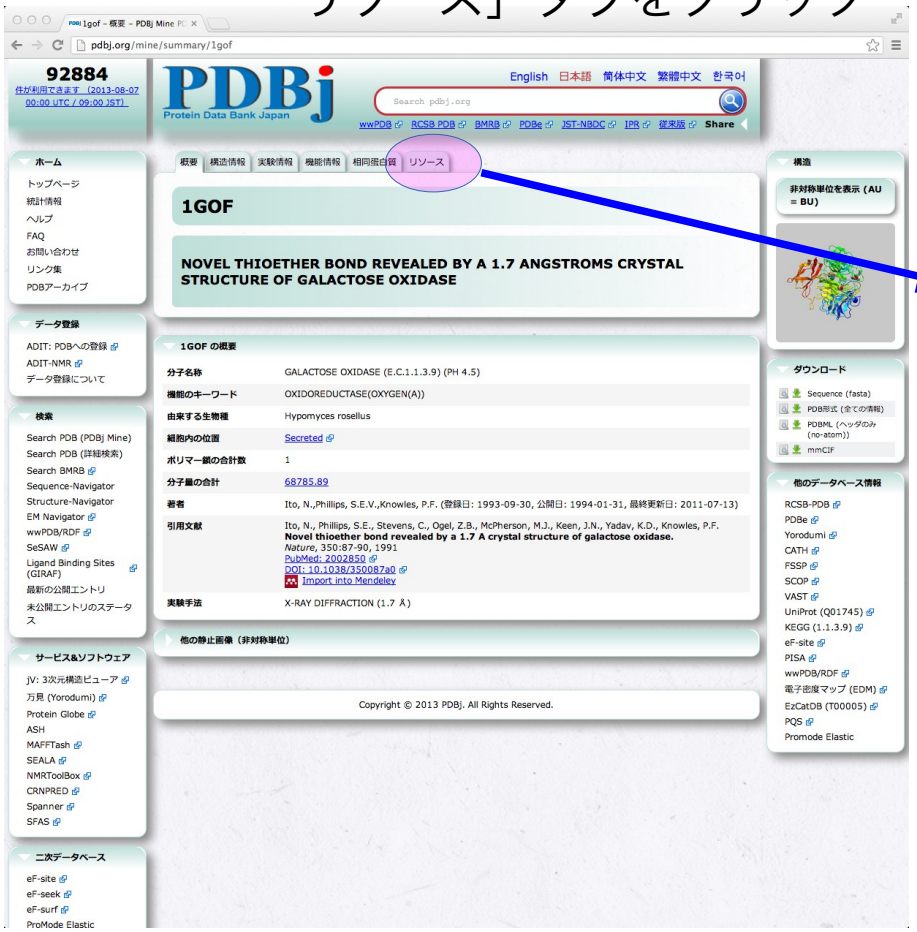
- XML を扱うソフトウェアが充実している。
  - 読み込み
  - 書き出し
  - 検証（XML スキーマを使う）
  - フォーマットの変換

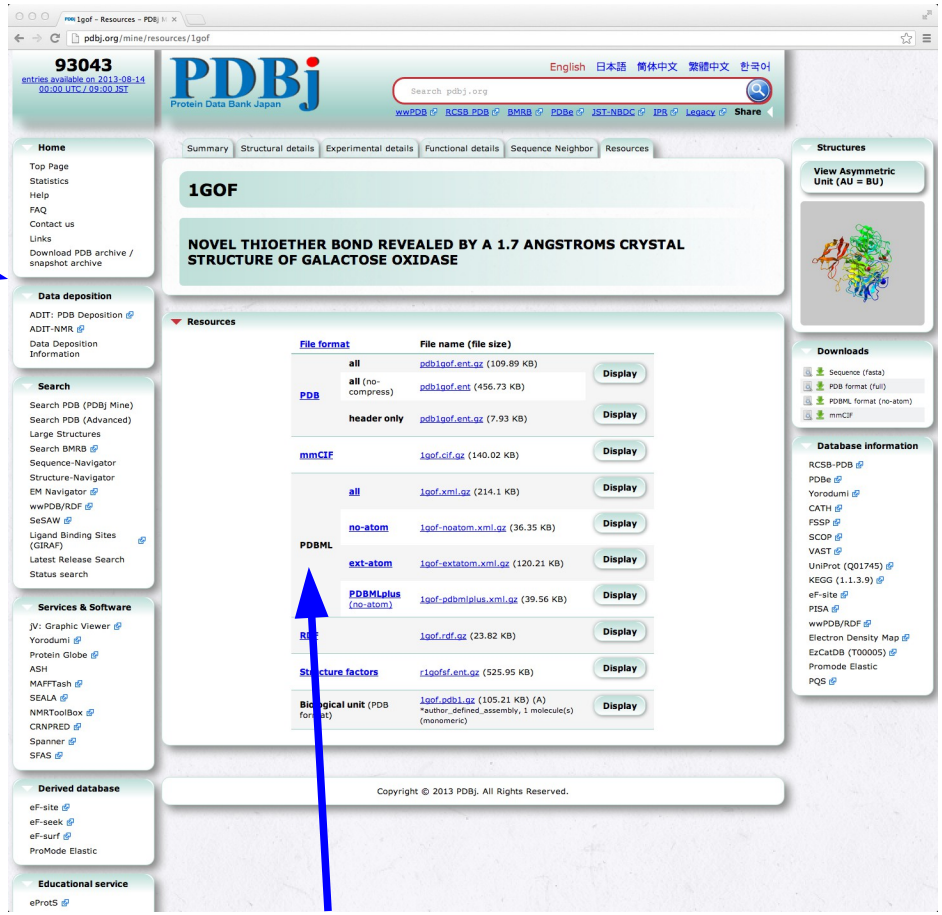
PDBML を使わない理由

- mmCIF を駆使できる場合
- PDB のデータを「目で」読みたい場合
- 大きなファイルサイズが負担になる場合

実際に PDBML を見てみる

「リソース」タブをクリック





「PDBML」に注目

PDBML ファイルの種類

- “all”
  - mmCIF に含まれる全ての情報
- “no-atom”
  - “all” から atom_site(原子座標) の情報を除いたもの
- “ext-atom”
  - atom_site のデータのみを簡略化して記述したもの
- “PDBMLplus”
  - “no-atom” のデータに PDBj が独自にアノテーションを加えたもの

ファイルサイズ

PDB エントリ 「1GOF」の場合

フォーマット	サイズ（バイト）	行数
PDB	457K	5774
mmCIF	595K	7925
PDBML(all)	5.6M	119719
PDBML(no-atom)	820K	16597
PDBML(extatom)	889K	5166

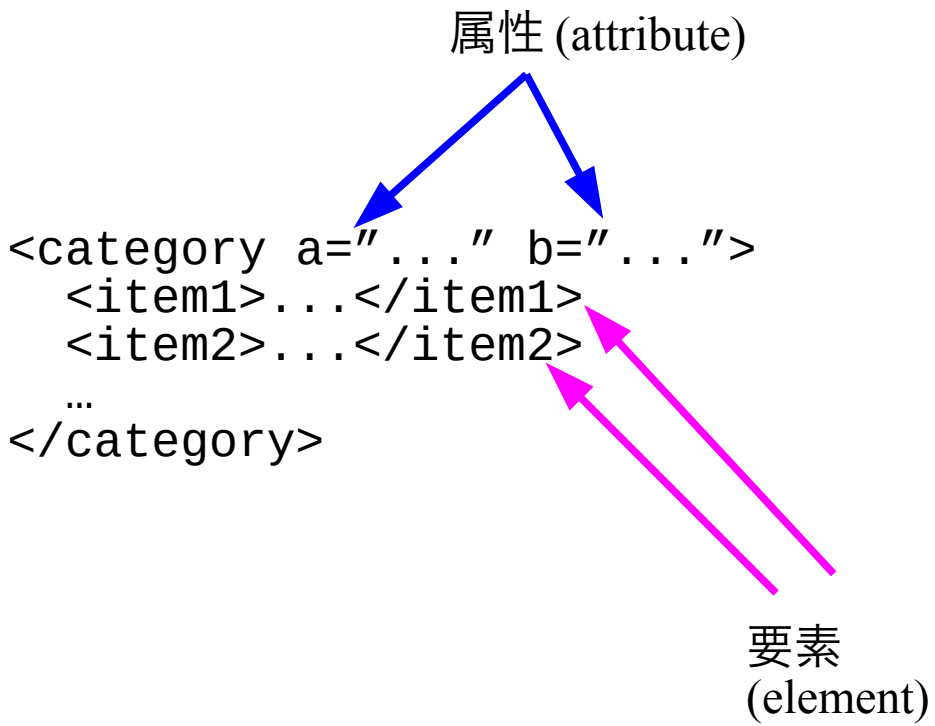
PDBML の基本構造

```
<?xml version="1.0" encoding="UTF-8" ?>

<PDBx:datablock datablockName="1GOF"
   xmlns:PDBx="http://pdbml.pdb.org/schema/pdbx-v40.xsd"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://pdbml.pdb.org/schema/pdbx-v40.xsd ...">

   <PDBx:atom_siteCategory>
      <PDBx:atom_site id="1">
         <PDBx:B_iso_or_equiv>34.65</PDBx:B_iso_or_equiv>
         <PDBx:Cartn_x>38.840</PDBx:Cartn_x>
         … ( そのカテゴリの項目が繰り返す )
      </PDBx:atom_site>
      <PDBx:atom_site id="2">
         <PDBx:B_iso_or_equiv>42.26</PDBx:B_iso_or_equiv>
         <PDBx:Cartn_x>38.356</PDBx:Cartn_x>
         …
      </PDBx:atom_site>
      … (atom_site が繰り返す )
   </PDBx:atom_siteCategory>
   … (xxxCategory が繰り返す )
</PDBx:datablock>
```

属性と要素の使い分け



- 各カテゴリの属性と要素はともにそのカテゴリの項目 (item)
- 属性はそのカテゴリの「主キー」に相当する

mmCIF と PDBML の比較

mmCIF

```
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Ito, N.'           1
'Phillips, S.E.V.' 2
'Knowles, P.F.'    3
```

PDBML

```
<PDBx:audit_authorCategory>
   <PDBx:audit_author pdbx_ordinal="1">
      <PDBx:name>Ito, N.</PDBx:name>
   </PDBx:audit_author>
   <PDBx:audit_author pdbx_ordinal="2">
      <PDBx:name>Phillips, S.E.V.</PDBx:name>
   </PDBx:audit_author>
   <PDBx:audit_author pdbx_ordinal="3">
      <PDBx:name>Knowles, P.F.</PDBx:name>
   </PDBx:audit_author>
</PDBx:audit_authorCategory>
```

atom_site(原子座標) をみてみる

```
<PDBx:atom_site id="1">
   <PDBx:B_iso_or_equiv>34.65</PDBx:B_iso_or_equiv>
   <PDBx:Cartn_x>38.840</PDBx:Cartn_x>
   <PDBx:Cartn_y>0.236</PDBx:Cartn_y>
   <PDBx:Cartn_z>1.012</PDBx:Cartn_z>
   <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
   <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
   <PDBx:auth_comp_id>ALA</PDBx:auth_comp_id>
   <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
   <PDBx:group_PDB>ATOM</PDBx:group_PDB>
   <PDBx:label_alt_id xsi:nil="true" />
   <PDBx:label_asym_id>A</PDBx:label_asym_id>
   <PDBx:label_atom_id>N</PDBx:label_atom_id>
   <PDBx:label_comp_id>ALA</PDBx:label_comp_id>
   <PDBx:label_entity_id>1</PDBx:label_entity_id>
   <PDBx:label_seq_id>1</PDBx:label_seq_id>
   <PDBx:occupancy>1.00</PDBx:occupancy>
   <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
   <PDBx:type_symbol>N</PDBx:type_symbol>
</PDBx:atom_site>
```

extatom

```
<PDBx:datablock datablockName="1GOF-extatom"
   xmlns:PDBx="http://pdbml.pdb.org/schema/pdbx-v40-ext.xsd"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://pdbml.pdb.org/schema/pdbx-v40-ext.xsd pdbx-v40-ext.xsd">

  <PDBx:category_atom_record>
    <PDBx:atom_record id="1">ATOM 1 A A 1 1 ? . ALA ALA N N  N  38.840  0.236 1.012 1.00 34.65 1 ?</PDBx:atom_record>
    <PDBx:atom_record id="2">ATOM 1 A A 1 1 ? . ALA ALA C CA CA 38.356 -0.999 0.357 1.00 42.26 1 ?</PDBx:atom_record>
    <PDBx:atom_record id="3">ATOM 1 A A 1 1 ? . ALA ALA C C  C  37.098 -1.547 1.056 1.00 41.25 1 ?</PDBx:atom_record>
    <PDBx:atom_record id="4">ATOM 1 A A 1 1 ? . ALA ALA O O  O  36.619 -0.946 2.028 1.00 29.44 1 ?</PDBx:atom_record>
      …...
```

- atom_site カテゴリのよく使う項目だけを一行で空白区切りで記述したもの
  → XML の精神には適合しないが、データ量は圧縮できる
- 記述の定義は http://pdbml.pdb.org/schema/pdbx-v40-ext.xsd を参照のこと

参考文献

- mmCIF
  - J.D. Westbrook, P.E. Bourne, *Bioinformatics* **16**:159-168 (2000)
  - http://mmcif.rcsb.org/
- PDBML
  - J.D. Westbrook, N. Ito, H. Nakamura, K. Henrick, H. M. Berman, *Bioinformatics* **21**:988-992 (2005)
  - http://pdbml.rcsb.org/

演習問題 (1)

- PDB エントリ　1GOF の
  - PDB ファイル
  - mmCIF ファイル
  - PDBML(all) ファイル

  をダウンロードしてみる。

演習問題（2）

- 1GOF の mmCIF ファイルで以下のカテゴリの内容を確認する
  - struct
  - entity
  - entity_poly
  - pdbx_entity_nonpoly
  - struct_asym
  - struct_ref

演習問題（3）

- 先ほど調べたカテゴリの各項目の意味を PDBx 辞書を使って調べてみる。
http://mmcif.pdb.org/dictionaries/mmcif_pdbx_v40.dic/Index/index.html

演習問題（4）

- カテゴリ間の関係を調べる
  - entity と struct_asym
  - entity と struct_ref
  - pdbx_entity_nonpoly と chem_comp
  - citation と citation_author

演習問題（5）

- カテゴリ間の関係は PDBx 辞書ではどのように記述されているか調べてみる。