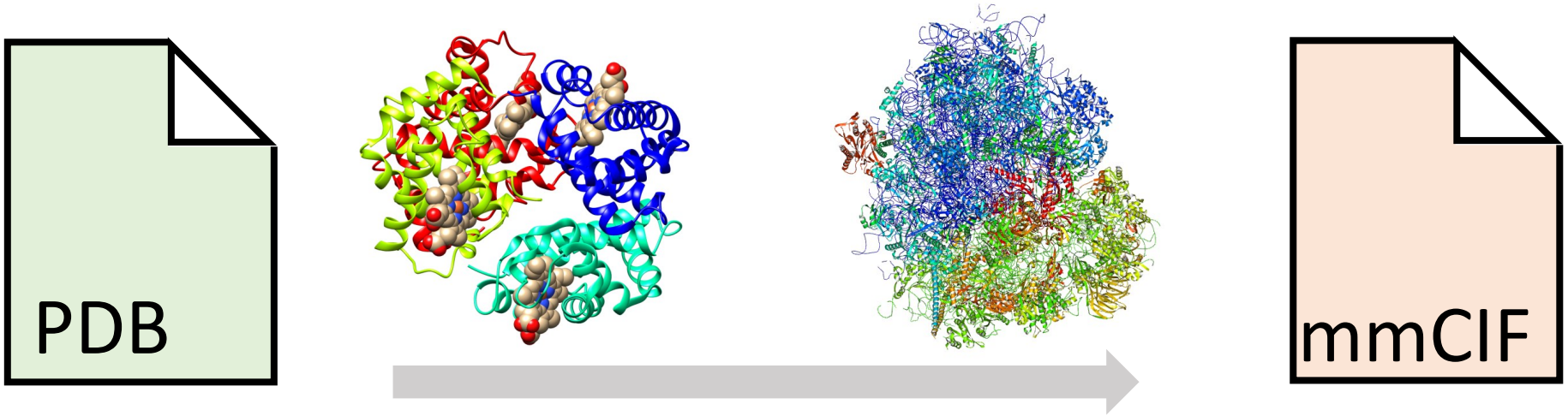
第５８回日本生物物理学会年会　オンライン展示会

旧PDBフォーマットと mmCIFフォーマット

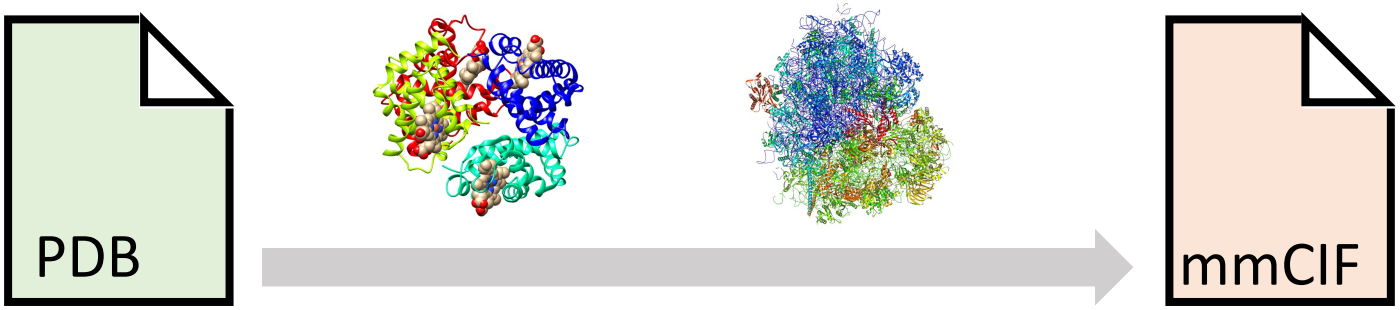


大阪大学　蛋白質研究所
蛋白質データベース開発研究室
特任准教授　川端　猛

2020年 9月16日(水)--18日(木)

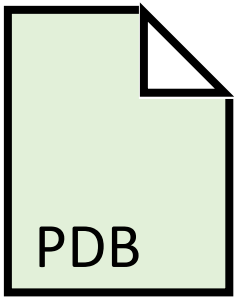
今日の話題

- 生体高分子の立体構造のデータは、これまで(旧)PDB形式で書かれていましたが、最近mmCIF形式に変わりました。
- 立体構造データのユーザー・プログラマーを対象に、mmCIFフォーマットの読み方、活用の仕方について説明します。
- データ登録用のmmCIFファイルの作成法については今回は説明しません。
- 立体構造をただ眺めるだけなら、ファイルフォーマットは意識する必要はありません。それでも、どんな情報が書かれているか概略は知っておいて損はないです。

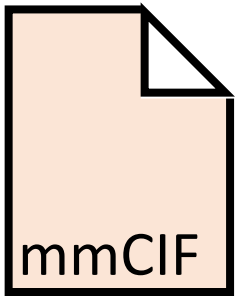


立体構造データのフォーマット

- 旧PDBフォーマット
  - PDB創立のころ(1971年)から使われていたフォーマット
  - １行１原子。１行８０文字の固定長のフォーマット。
  - 鎖識別子は１文字なので、62鎖を超えると扱えない。
  - 原子番号は５文字なので、99999個を超える原子は扱えない。
  - 座標以外のデータの記述は場当たり的



- mmCIFフォーマット
  - STARフォーマットの一種。可変長。
  - キー・バリュー形式と表形式を併用
  - 値の長さは可変なので、鎖や原子数の上限はない。
  - 座標以外の様々なデータも、柔軟に記述できる
  - 2014年から、PDBの正規なフォーマットとなる
  - 2019年7月から、 X線のデータ登録にmmCIFが必須となる
  - 主要な分子表示、モデリング、構造決定のソフトウエアは対応済み

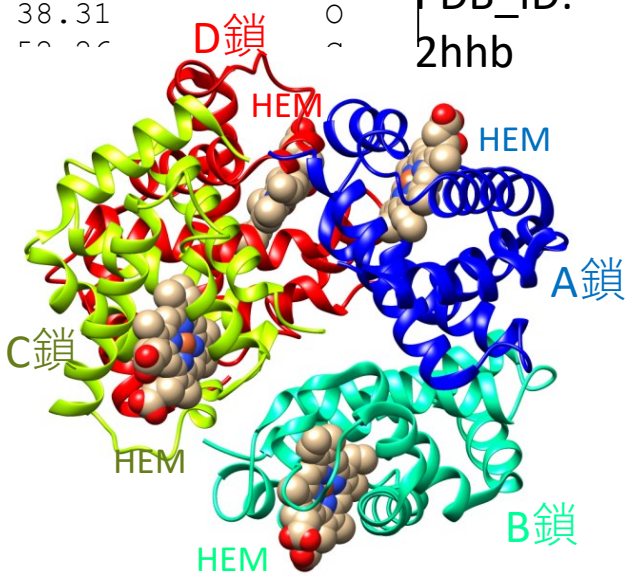


旧PDBフォーマットの立体構造データの例

PDB

PDB_ID: 2hhb

```
HEADER     OXYGEN TRANSPORT                          07-MAR-84   2HHB
TITLE      THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS
ATOM       1  N   VAL A   1        6.130  16.559   4.905  1.00 41.29           N
ATOM       2  CA  VAL A   1        6.870  17.784   4.702  1.00 41.33           C
ATOM       3  C   VAL A   1        8.377  17.548   4.913  1.00 31.64           C
ATOM       4  O   VAL A   1        8.820  16.980   5.922  1.00 38.31           O
ATOM       5  CB  VAL A   1        6.345  18.763   5.731  1.00
ATOM       6  CG1 VAL A   1        6.745  20.188   5.356  1.00
ATOM       7  CG2 VAL A   1        4.826  18.612   5.847  1.00
ATOM       8  N   LEU A   2        9.146  18.005   3.962  1.00
ATOM       9  CA  LEU A   2       10.599  17.914   4.153  1.00
ATOM      10  C   LEU A   2       11.062  19.085   5.062  1.00
ATOM      11  O   LEU A   2       10.829  20.254   4.744  1.00
ATOM      12  CB  LEU A   2       11.269  18.078   2.776  1.00
ATOM      13  CG  LEU A   2       10.986  16.983   1.769  1.00
ATOM      14  CD1 LEU A   2       11.735  17.168   0.427  1.00
ATOM      15  CD2 LEU A   2       11.276  15.630   2.404  1.00
ATOM      16  N   SER A   3       11.674  18.766   6.158  1.00
ATOM      17  CA  SER A   3       12.286  19.774   7.034  1.00
ATOM      18  C   SER A   3       13.529  20.322   6.334  1.00
ATOM      19  O   SER A   3       13.995  19.754   5.344  1.00 27.40           O
ATOM      20  CB  SER A   3       12.719  19.060   8.326  1.00 23.83           C
ATOM      21  OG  SER A   3       13.844  18.245   8.107  1.00 29.07           O
:
HETATM 4389  CHA HEM A 142        8.634   7.898 -18.334  1.00 17.24           C
HETATM 4390  CHB HEM A 142       10.386   9.965 -14.276  1.00 23.17           C
HETATM 4391  CHC HEM A 142        8.337   6.403 -11.611  1.00 11.04           C
:    原子番号 原子名 残基名  残基番号    X座標     Y座標     Z座標    占有率 温度因子      元素名
                          鎖識別子
```



80文字の固定長

パンチカード時代の名残り

mmCIFフォーマットの立体構造データの例

mmCIF

キー・バリュー形式

```
data_2HHB
#
_entry.id   2HHB
#
_struct.title                    'THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS
_struct.pdbx_descriptor          'HEMOGLOBIN (DEOXY)'
#
```

表形式

項目は**[カテゴリ名].[アイテム名]**

※2hhbの場合58個のカテゴリがある

```
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.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   . VAL A 1 1   ? 6.130   16.559  4.905   1.00 41.29 ? 1   VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? 6.870   17.784  4.702   1.00 41.33 ? 1   VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? 8.377   17.548  4.913   1.00 31.64 ? 1   VAL A C   1
ATOM   4    O  O   . VAL A 1 1   ? 8.820   16.980  5.922   1.00 38.31 ? 1   VAL A O   1
ATOM   5    C  CB  . VAL A 1 1   ? 6.345   18.763  5.731   1.00 52.26 ? 1   VAL A CB  1
ATOM   6    C  CG1 . VAL A 1 1   ? 6.745   20.188  5.356   1.00 52.75 ? 1   VAL A CG1 1
ATOM   7    C  CG2 . VAL A 1 1   ? 4.826   18.612  5.847   1.00 58.75 ? 1   VAL A CG2 1
ATOM   8    N  N   . LEU A 1 2   ? 9.146   18.005  3.962   1.00 27.63 ? 2   LEU A N   1
ATOM   9    C  CA  . LEU A 1 2   ? 10.599  17.914  4.153   1.00 33.62 ? 2   LEU A CA  1
:
HETATM 4385 C  CHA . HEM E 3 .   ? 8.634   7.898   -18.334 1.00 17.24 ? 142 HEM A CHA 1
HETATM 4386 C  CHB . HEM E 3 .   ? 10.386  9.965   -14.276 1.00 23.17 ? 142 HEM A CHB 1  A CHC 1
:
#
```

データベース管理者が入力(label_のアイテム)

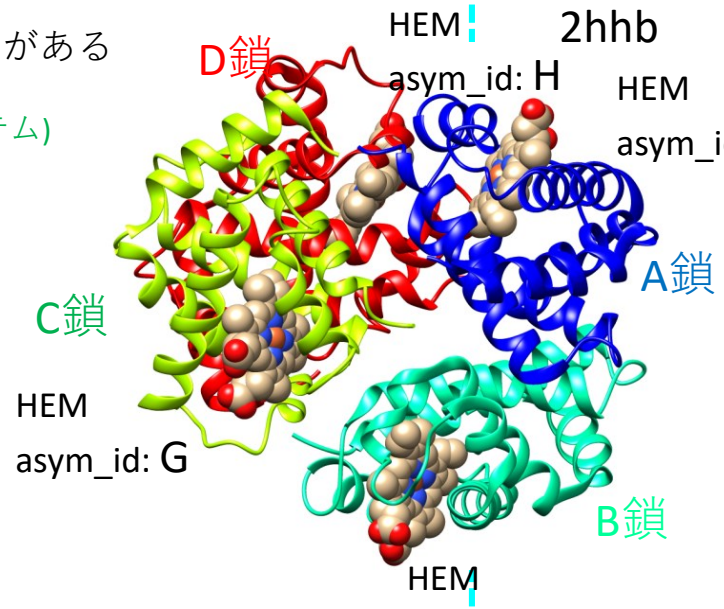
原子座標

登録者が入力する(auth_のアイテム)

原子番号　元素名　原子名　残基名　残基番号　asym_id(鎖識別子)　X座標　Y座標　Z座標　占有率　温度因子　残基番号　残基名　asym_id(鎖識別子)　原子名

entity_id

PDB_ID
2hhb



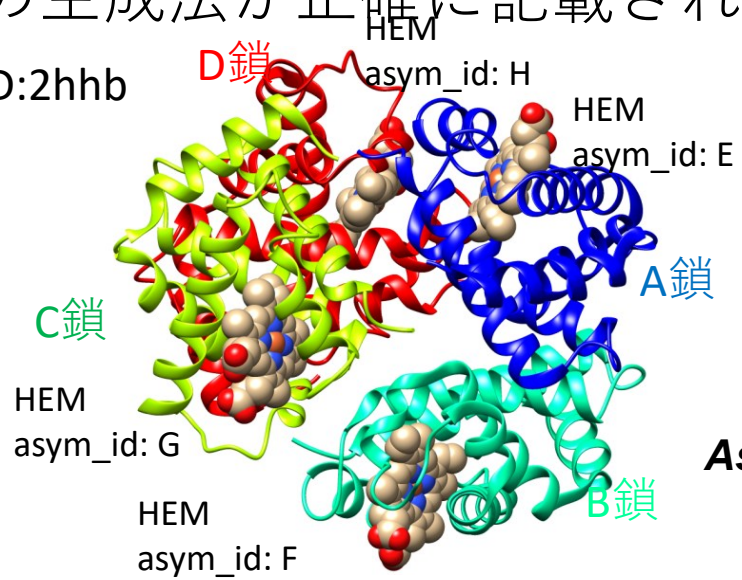
mmCIFファイルの特長

- STARフォーマット。キーバリュー形式と表形式の併用で、人間にも比較的読みやすい。
- **[カテゴリ名].[アイテム名]** で整理された項目名

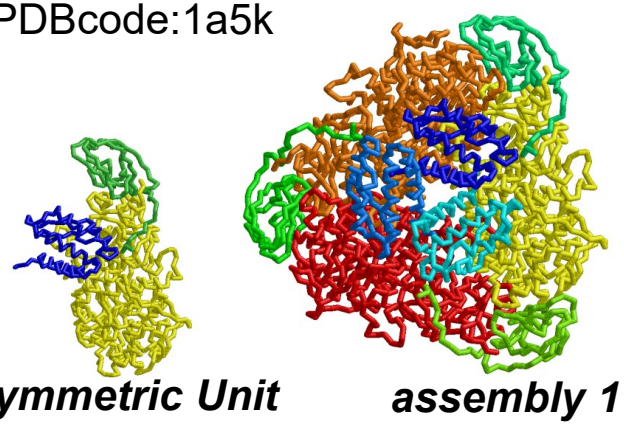
Relionの出力ファイルもSTARフォーマットです！

- カテゴリatom_site以外に、50-80ほどのカテゴリ（表）
- 鎖識別子、原子番号などの文字数は無制限
- 残基番号、鎖識別子については、読みやすさと統一性を両立するため、<u>著者入力</u>と<u>データベース管理者入力</u>の２種が書かれる。
- Biological unitなど、非対称単位以外の複合体(assembly)の生成法が正確に記載されている。

PDB_ID:2hhb



PDBcode:1a5k



mmCIFを使わなければならない場合

１．巨大構造を扱う場合は、mmCIFを読み込むべき

巨大な構造の場合、mmCIF形式のみで公開されています。

鎖識別子が２文字以上の場合、巨大でなくてもmmCIFだけの構造はあります

２．立体構造を扱うプログラムの開発者は、入力をmmCIFに対応するよう改良すべき。（出力は任意）

計算結果の出力は、旧PDBフォーマットでもかまいません。

３．PDBデータベース全体に対して、統計的・包括的な解析を行う場合は、mmCIFをベースにすべき

mmCIFのデータベース管理者による残基番号、鎖識別子は、
統一的な基準で決まっているので、PDBデータベース全体の統計を
とる場合に向いています。

巨大構造エントリ = mmCIFだけで公開されているエントリ

鎖識別子(auth_asym_id)を２文字以上使うと、原子数が小さくてもmmCIFだけの公開となり、「巨大構造エントリー」とみなされる。

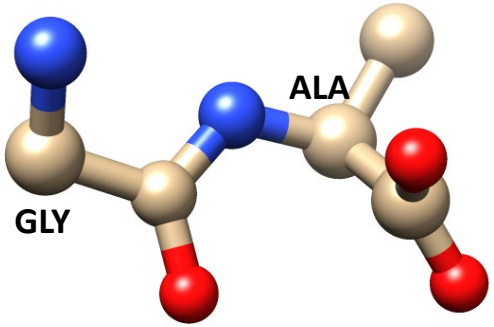
分子ビューアで表示できる最小のmmCIFファイル

モデル構造の出力にmmCIF形式を使いたい場合

```
data_1XXX
#
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.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   . GLY A 1 1 ? -1.539 -2.707 0.094  1.00 0.00 ? 1 GLY A N   1
ATOM 2  C CA  . GLY A 1 1 ? -0.090 -2.707 0.094  1.00 0.00 ? 1 GLY A CA  1
ATOM 3  C C   . GLY A 1 1 ? 0.433  -1.277 0.094  1.00 0.00 ? 1 GLY A C   1
ATOM 4  O O   . GLY A 1 1 ? 1.326  -0.942 0.869  1.00 0.00 ? 1 GLY A O   1
ATOM 5  N N   . ALA A 1 2 ? -0.126 -0.436 -0.779 1.00 0.00 ? 2 ALA A N   1
ATOM 6  C CA  . ALA A 1 2 ? 0.284  0.950  -0.877 1.00 0.00 ? 2 ALA A CA  1
ATOM 7  C C   . ALA A 1 2 ? 0.130  1.638  0.472  1.00 0.00 ? 2 ALA A C   1
ATOM 8  O O   . ALA A 1 2 ? 1.075  2.319  0.862  1.00 0.00 ? 2 ALA A O   1
ATOM 9  C CB  . ALA A 1 2 ? -0.563 1.702  -1.898 1.00 0.00 ? 2 ALA A CB  1
ATOM 10 O OXT . ALA A 1 2 ? -0.929 1.460  1.071  1.00 0.00 ? 2 ALA A OXT 1
#
```

(1) data_[PDB_ID]の行

(2) atom_siteカテゴリの表



これだけあれば、
PyMOL, RasMol,
UCSF Chimera
で構造は表示できる！

ただし、PDBへ登録用のmmCIFファイルはこれだけでは不十分なので要注意

なんとなく構造を眺めるだけなら…

- PDBのウェブページをクリックして、３D構造を眺めるだけなら、ファイルフォーマットを意識する必要はありません
- PyMOL、UCSF Chimeraなど最近のビューワは、直接PDBサーバから構造を読みにいく機能(“fetch”)があるので、ファイルフォーマットを意識する必要はありません
- ただし、残基名、残基番号、鎖識別子などの事情はある程度理解しておいたほうがいいかもしれません

→次に、２種の鎖識別子 (auth_asym_id, label_asym_id)とエンティティ(entity)、２種の残基番号(auth_seq_idとlabel_seq_id)の説明をします。

鎖識別子 (Chain ID)に関して

mmCIFには、鎖識別子(ChainID)という項目はなく、以下の三つのカテゴリで表現される

auth_asym_id :

従来のPDBのChain IDに相当。ただし、二文字以上でもかまわない。登録者が命名。低分子化合物は、結合している高分子と同じIDを振るのが普通。

label_asym_id :

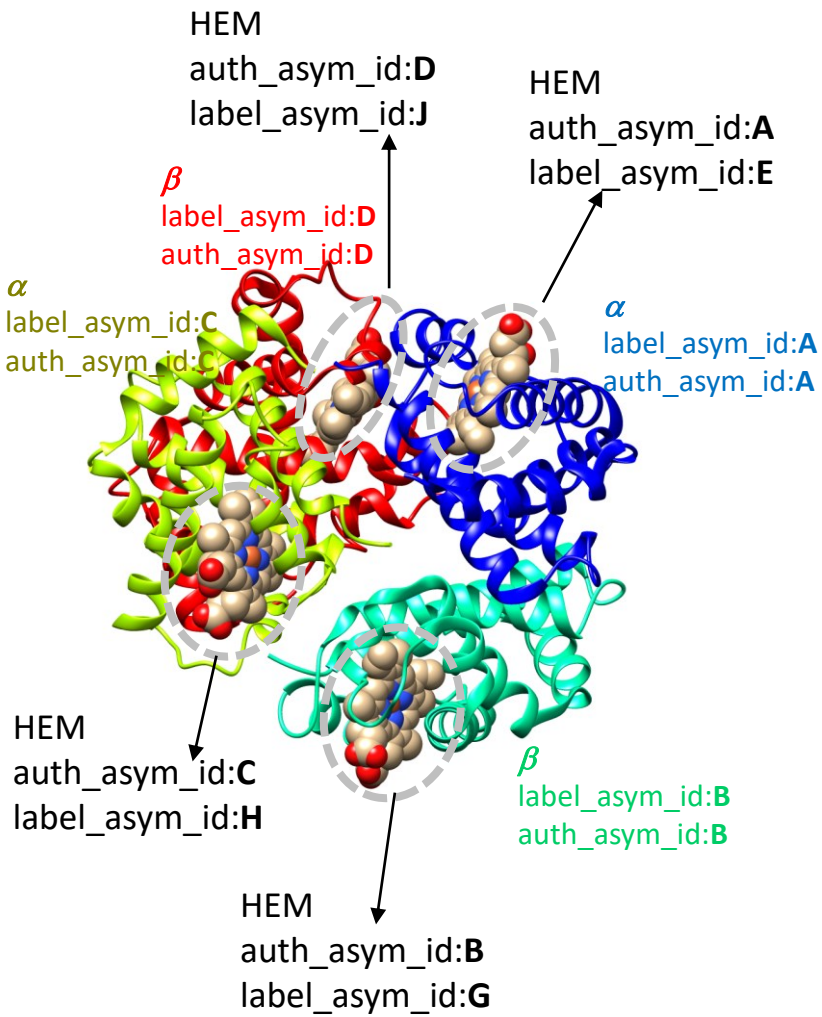
データベース管理者が入力するID。auth_asym_idと似ているが、原則的に一つの分子には固有のidを振る。低分子化合物も固有のIDを持つことになる。

entity :

分子の種類を示す番号。同種の分子が複数ある場合（ホモ多量体など）では、各分子は別のlabel_asym_idを持つが、entity番号は同一。UniProtのリンクはentityにつく。

ヘモグロビン(2hhb)の場合

auth_asym_id	label_asym_id	entity	comp_id	Pdbx_description
A	A	1 α		HEMOGLOBIN (DEOXY) (ALPHA CHAIN)'
B	B	2 β		HEMOGLOBIN (DEOXY) (BETA CHAIN)'
C	C	1 α		HEMOGLOBIN (DEOXY) (ALPHA CHAIN)'
D	D	2 β		HEMOGLOBIN (DEOXY) (BETA CHAIN)'
A	**E**	**3** HEM	**HEM**	**PROTOPORPHYRIN IX CONTAINING FE**
B	F	4	PO4	PHOSPHATE ION
B	**G**	**3** HEM	**HEM**	**PROTOPORPHYRIN IX CONTAINING FE**
C	**H**	**3** HEM	**HEM**	**PROTOPORPHYRIN IX CONTAINING FE**
D	I	4	PO4	PHOSPHATE ION
D	**J**	**3** HEM	**HEM**	**PROTOPORPHYRIN IX CONTAINING FE**



・４つのHEM分子はそれぞれ固有のlabel_asym_id

・entityはA鎖とC鎖は1, B鎖とD鎖は2.　α鎖とβ鎖に対応

PDBjのWEBページでの表示

各PDBエントリのページの[構造情報]のタブを選ぶと、Entityの表が表示されます。

概要 | 構造情報 | 実験情報 | 機能情報 | 相同蛋白質 | 履歴 | ダウンロード

2HHB

THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS RESOLUTION

「1HHB」から置き換えられました

エンティティ

entity　auth_asym_id

Entity ID	鎖名	説明	種類	鎖長	分子量	分子数	データベース名(アクセス番号)	由来する生物種	エンティティの一般名
1	A, C	HEMOGLOBIN (DEOXY) (ALPHA CHAIN)	polymer	141	15150.4	2	UniProt (P01922) UniProt (by SIFTS) (P69905)	Homo sapiens (human) @PDBj	
2	B, D	HEMOGLOBIN (DEOXY) (BETA CHAIN)	polymer	146	15890.2	2	UniProt (P02023) UniProt (by SIFTS) (P68871)	Homo sapiens (human) @PDBj	
3	A, B, C, D	PROTOPORPHYRIN IX CONTAINING FE	non-polymer		616.5	4	Chemie (HEM)		
4	B, D	PHOSPHATE ION	non-polymer		95.0	2	Chemie (PO4)		
5		water	water		18.0	221	Chemie (HOH)		

蛋白質の場合、アミノ酸配列のID (UniProt AC)は、entityと１対１対応する

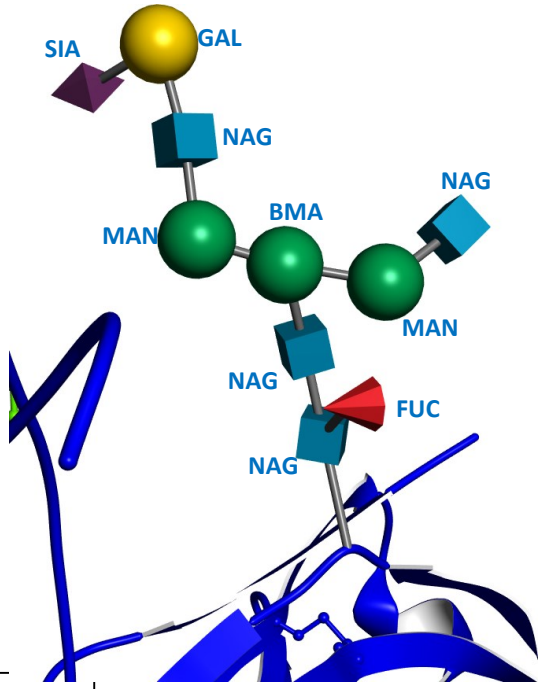
Carbohydrate remediation (2020/07/29) によるentityの変更の例

PDB_ID:4fqc 旧データ (2020/07/22)

```
loop_
_entity.id  _entity.type  _entity.src_method  _entity.pdbx_description
_entity.formula_weight  _entity.pdbx_number_of_molecules  _entity.pdbx_ec
_entity.pdbx_mutation  _entity.pdbx_fragment  _entity.details
1  polymer     man 'Fab heavy chain'                    26343.484 1   ? ? ? ?
2  polymer     man 'Fab light chain'                    22602.961 1   ? ? ? ?
3  non-polymer man BETA-D-MANNOSE   BMA                 180.156   1   ? ? ? ?
4  non-polymer man ALPHA-D-MANNOSE  MAN                 180.156   2   ? ? ? ?
5  non-polymer man N-ACETYL-D-GLUCOSAMINE NAG           221.208   4   ? ? ? ?
6  non-polymer man BETA-D-GALACTOSE GAL                 180.156   1   ? ? ? ?
7  non-polymer man ALPHA-L-FUCOSE FUC                   164.156   1   ? ? ? ?
8  non-polymer man 'O-SIALIC ACID' SIA                  309.270   1   ? ? ? ?
9  non-polymer syn 'TRIS(HYDROXYETHYL)AMINOMETHANE' 163.215   1   ? ? ? ?
10 water       nat water                                18.015    146 ? ? ? ?
```

PDB_ID:4fqc 新データ (2020/07/29)

```
loop_
_entity.id  _entity.type_entity.src_method  _entity.pdbx_description
_entity.formula_weight  _entity.pdbx_number_of_molecules
_entity.pdbx_ec  _entity.pdbx_mutation  _entity.pdbx_fragment  _entity.details
1 polymer     man 'Fab heavy chain' 26343.484 1   ? ? ? ?
2 polymer     man 'Fab light chain' 22602.961 1   ? ? ? ?
3 branched    man
;N-acetyl-alpha-neuraminic acid-(2-6)-beta-D-galactopyranose-(1-4)-2-acetamido-
2-deoxy-beta-D-glucopyranose-(1-2)-alpha-D-mannopyranose-(1-3)-[2-acetamido-2-
deoxy-beta-D-glucopyranose-(1-2)-alpha-D-mannopyranose-(1-6)]beta-D-
mannopyranose-(1-4)-2-acetamido-2-deoxy-beta-D-glucopyranose-(1-4)-[alpha-L-
fucopyranose-(1-6)]2-acetamido-2-deoxy-beta-D-glucopyranose
;
1916.745  1   ? ? ? ?
4 non-polymer syn 'TRIS(HYDROXYETHYL)AMINOMETHANE' 163.215   1   ? ? ? ?
5 water       nat water 18.015    146 ? ? ? ?
```



PDB_ID:4fqc
ASNに共有結合している糖鎖

旧データ：糖鎖を構成する各単糖が別のentityを構成

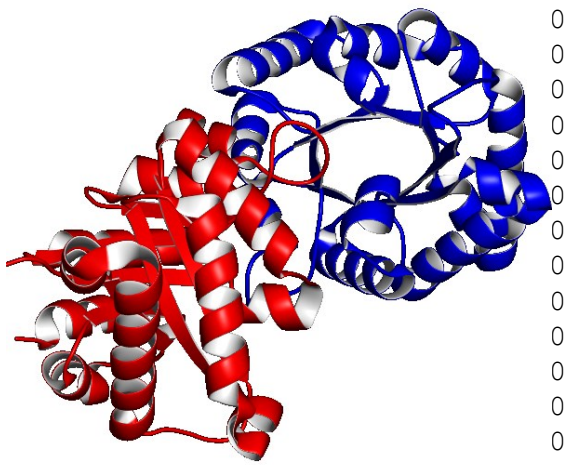
新データ：糖鎖全体が一つの”branched”という型のentity.

残基番号について(label_seq_id, auth_seq_id)

PDB_ID:1tim　STRUCTURE OF TRIOSE PHOSPHATE ISOMERASE FROM CHICKEN MUSCLE

```
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.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   ? 43.240 11.990 -6.915  1.00 0.00 ? 1   ALA A N   1
ATOM 2    C CA  . ALA A 1 1   ? 43.888 10.862 -6.231  1.00 0.00 ? 1   ALA A CA  1
ATOM 3    C C   . ALA A 1 1   ? 44.791 11.378 -5.094  1.00 0.00 ? 1   ALA A C   1
ATOM 4    O O   . ALA A 1 1   ?                            0 ? 1   ALA A O   1
ATOM 5    C CB  . ALA A 1 1   ?                            0 ? 1   ALA A CB  1
ATOM 6    N N   . PRO A 1 2   ?                            0 ? 2   PRO A N   1
ATOM 7    C CA  . PRO A 1 2   ?                            0 ? 2   PRO A CA  1
ATOM 8    C C   . PRO A 1 2   ?                            0 ? 2   PRO A C   1
ATOM 9    O O   . PRO A 1 2   ?                            0 ? 2   PRO A O   1
ATOM 10   C CB  . PRO A 1 2   ?                            0 ? 2   PRO A CB  1
ATOM 11   C CG  . PRO A 1 2   ?                            0 ? 2   PRO A CG  1
ATOM 12   C CD  . PRO A 1 2   ?                            0 ? 2   PRO A CD  1
ATOM 13   N N   . ARG A 1 3   ?                            0 ? 4   ARG A N   1
ATOM 14   C CA  . ARG A 1 3   ?                            0 ? 4   ARG A CA  1
ATOM 15   C C   . ARG A 1 3   ?                            0 ? 4   ARG A C   1
ATOM 16   O O   . ARG A 1 3   ?                            0 ? 4   ARG A O   1
ATOM 17   C CB  . ARG A 1 3   ?                            0 ? 4   ARG A CB  1
ATOM 18   C CG  . ARG A 1 3   ?                            0 ? 4   ARG A CG  1
ATOM 19   C CD  . ARG A 1 3   ?                            0 ? 4   ARG A CD  1
ATOM 20   N NE  . ARG A 1 3   ?                            0 ? 4   ARG A NE  1
ATOM 21   C CZ  . ARG A 1 3                                4   ARG A CZ  1
ATOM 22   N NH1 . ARG A 1 3                                4   ARG A NH1 1
ATOM 23   N NH2 . ARG A 1 3                                4   ARG A NH2 1
ATOM 24   N N   . LYS A 1 4   ? 47.060 16.521 -2.977  1.00 0.00 ? 5   LYS A N   1
```

label_seq_id　　auth_seq_id



※auth_seq_idで3を飛ばしたのはヒトの残基番号に合わせるため

```
loop_
_entity_poly_seq.entity_id
_entity_poly_seq.num
_entity_poly_seq.mon_id
_entity_poly_seq.hetero
1 1   ALA n
1 2   PRO n
1 3   ARG n
1 4   LYS n
1 5   PHE n
1 6   PHE n
1 7   VAL n
1 8   GLY n
1 9   GLY n
1 10  ASN n
:
```

auth_seq_idは
1,2,4,5,6
label_seq_idは
1,2,3,4,5

label_seq_idを残基番号として使ったほうが、計算機による処理に向いている。

mmCIFファイルを読み込む方法

1．旧PDB形式に変換して読み込む

a) PDBで公開されている変換済のファイルを使う

b)変換プログラムを使う

2．自分で読み込むプログラムを作成する

a) STAR形式を読み込むプログラムを自作する

b) atom_siteの表だけを読むようにする

3．ライブラリを使用する

PDBで公開されている旧PDBフォーマットのファイル

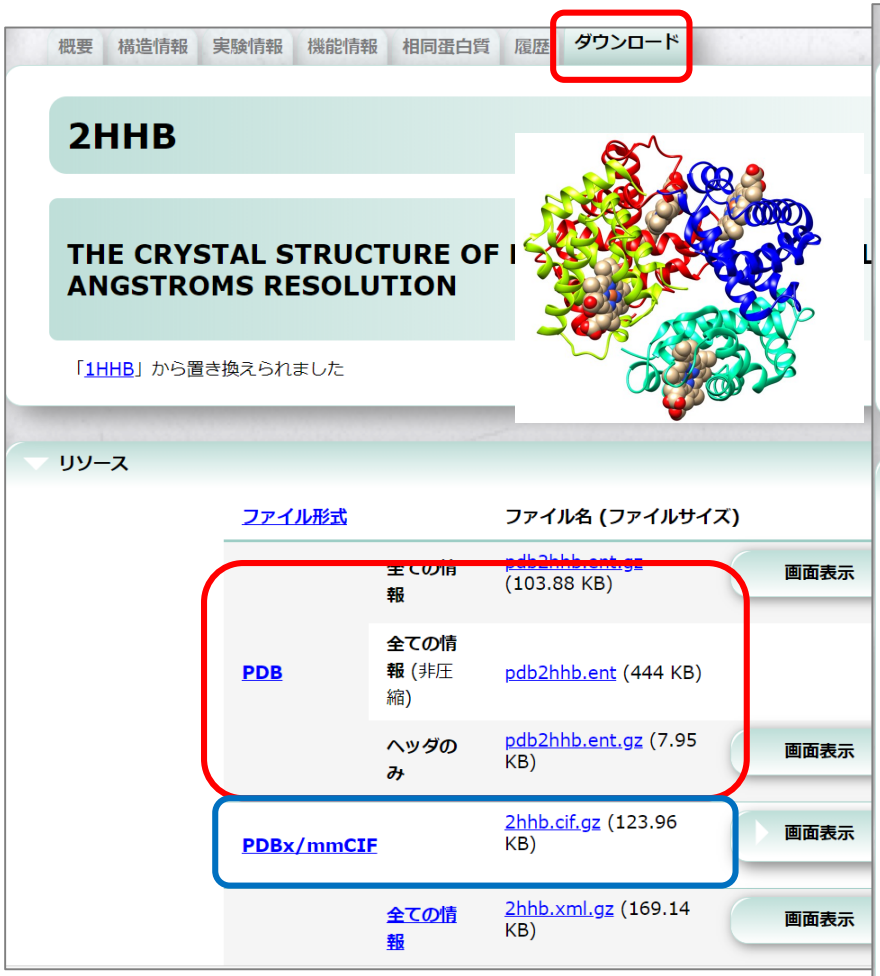
1. 通常のエントリは、pdb[PDB_ID].entのファイル名でFTPサイト等で公開されている

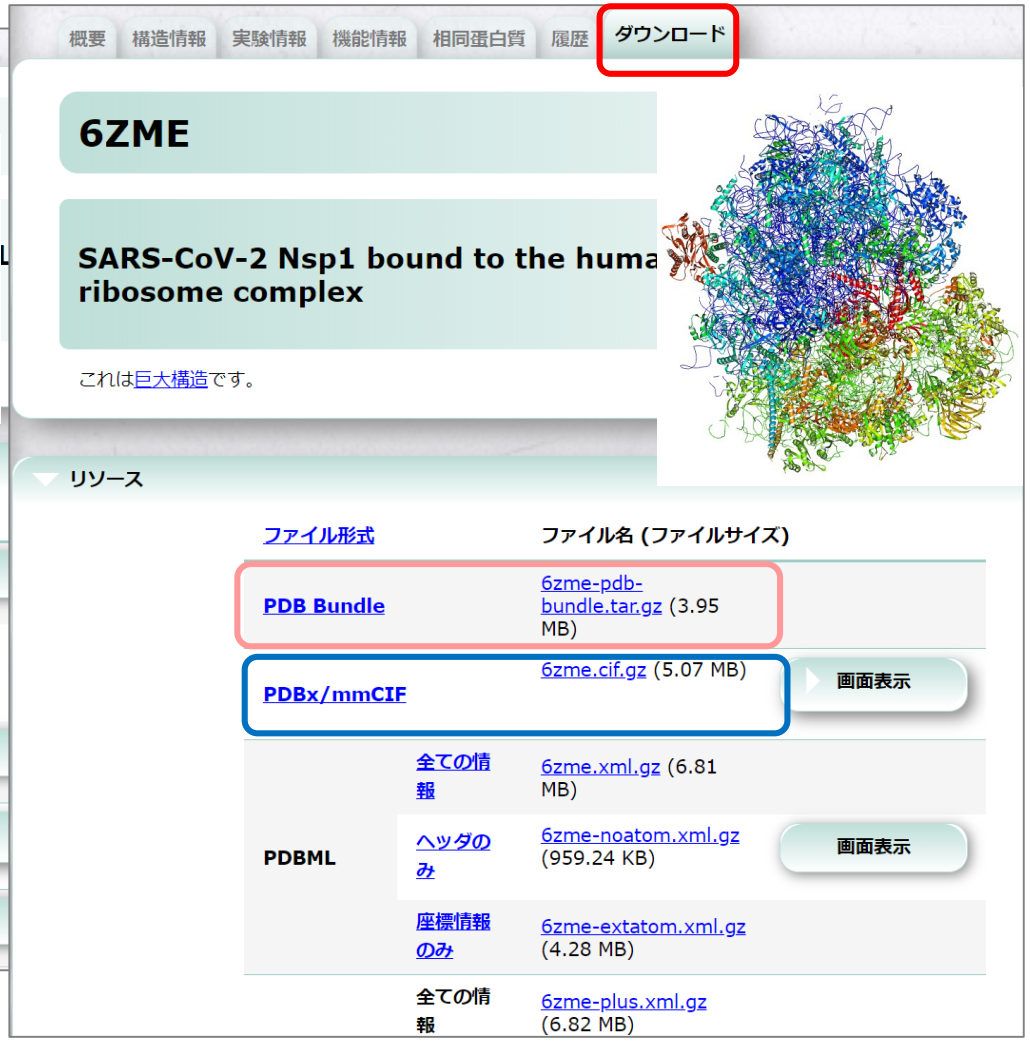
2. 「巨大構造」エントリは旧PDBフォーマットのファイルはない。救済として、一つのエントリを複数のPDBファイルに分割して表現されたbundleファイル が公開されている。

例えば、PDB_ID:6zmeの場合、
6zme-pdb-bundle.tar.gzというファイルが公開されている。これを展開すると、

```
6zme-pdb-bundle1.pdb
6zme-pdb-bundle2.pdb
6zme-pdb-bundle3.pdb
6zme-chain-id-mapping.txt
```

PDBjの[ダウンロード]タブ

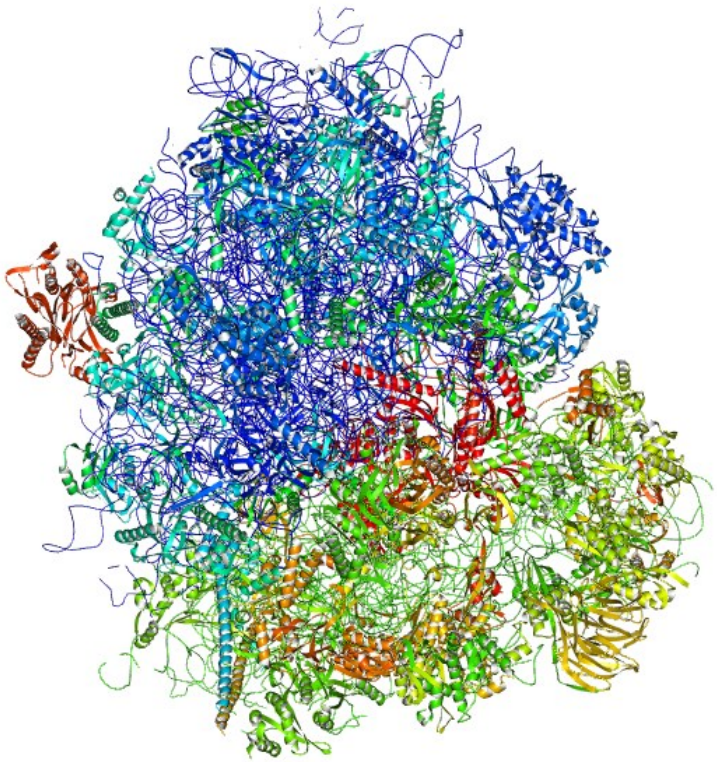




bundleファイルの例

例えば、PDB_ID:6zmeの場合、6zme-pdb-bundle.tar.gzというファイルがPDBで公開されている。これを展開すると、以下の三つのPDBファイルができる。

6zme-pdb-bundle1.pdb
6zme-pdb-bundle2.pdb
6zme-pdb-bundle3.pdb
6zme-chain-id-mapping.txt



6zme-chain-id-mapping.txt
PDBの鎖名とmmCIFの鎖名の対応表

```
 New chain ID          Original chain ID
6zme-pdb-bundle1.pdb:
            A                      L5
            B                      L7
:
            H                      LE

6zme-pdb-bundle2.pdb:
            A                      LF
            B                      LG
:
            0                      SR

6zme-pdb-bundle3.pdb:
            A                      SS
            B                      ST
:
            a                      CH
```

mmCIF⇔PDB変換プログラム

Web上で使える変換サービスが公開されています。

https://mmcif.pdbj.org/converter/index.php?l=ja

mmCIF→PDB変換プログラム

例えばmaxitというプログラムの場合

```
maxit -input 6zme.cif -output 6zme.pdb -o 2
```

mmCIFから変換されたPDBファイルの例

```
ATOM      1  O5'   CL5   1      158.480 257.280 294.058  1.00 77.16          O
ATOM      2  C5'   CL5   1      159.390 256.413 293.392  1.00 77.16          C
ATOM      3  C4'   CL5   1      159.635 256.854 291.971  1.00 77.16          C
ATOM      4  O4'   CL5   1      160.046 258.247 291.971  1.00 77.16          O
ATOM      5  C3'   CL5   1      160.727 256.089 291.228  1.00 77.16          C
ATOM      6  O3'   CL5   1      160.399 256.025 289.842  1.00 77.16          O
ATOM      7  C2'   CL5   1      161.945 256.984 291.415  1.00 77.16          C
ATOM      8  O2'   CL5   1      162.948 256.820 290.435  1.00 77.16          O
ATOM      9  C1'   CL5   1      161.312 258.372 291.354  1.00 77.16          C
ATOM     10  N1    CL5   1      162.066 259.405 292.075  1.00 77.16          N
:
ATOM  99999  O   GLNCH 420      275.701 168.452 172.461  1.00 55.68          O
ATOM  99999  CB  GLNCH 420      274.539 168.241 169.904  1.00 55.68          C
ATOM  99999  CG  GLNCH 420      273.850 168.776 168.660  1.00 55.68          C
ATOM  99999  CD  GLNCH 420      274.267 170.196 168.327  1.00 55.68          C
ATOM  99999  OE1 GLNCH 420      275.264 170.702 168.842  1.00 55.68          O
ATOM  99999  NE2 GLNCH 420      273.503 170.847 167.459  1.00 55.68          N
ATOM  99999  N   GLYCH 421      273.869 167.648 173.487  1.00 50.36          N
ATOM  99999  CA  GLYCH 421      274.532 167.463 174.765  1.00 50.36          C
ATOM  99999  C   GLYCH 421      274.287 168.592 175.748  1.00 50.36          C
ATOM  99999  O   GLYCH 421      275.012 168.738 176.732  1.00 50.36          O
```

原子番号は99999以降
増加させない

２文字の鎖名を表現するために、残基名と鎖名の間の列を使う。

mmCIFファイルを読み込むプログラムの作成方針

1．mmCIFファイルを多数行からなる文字列として読み込む

2．文字列の行を複数の「表」として読み込む（キー・バリュー形式の部分も表として読む）

3．「表」から必要な項目を取り出し「分子構造」のデータとする

mmCIFフォーマットを「表」として読む

```
#
loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
1 polymer     nat 'HEMOGLOBIN (DEOXY) (ALPHA CHAIN)' 15150.353 2
2 polymer     nat 'HEMOGLOBIN (DEOXY) (BETA CHAIN)'  15890.198 2
3 non-polymer syn 'PROTOPORPHYRIN IX CONTAINING FE'  616.487   4
4 non-polymer syn 'PHOSPHATE ION'                    94.971    2
5 water       nat water                              18.015
221
#
```

Pythonの場合：

D[表名][列名][行番号]

```
D['entity']['type'][0]='polymer'
D['entity']['type'][1]='polymer'
D['entity']['type'][2]='non-polymer'
:
```

entity

id	type	src_method	pdbx_description	formula_weight	pdbx_number_of_molecules
1	polymer	nat	'HEMOGLOBIN (DEOXY) (ALPHA CHAIN)'	15150.353	2
2	polymer	nat	'HEMOGLOBIN (DEOXY) (BETA CHAIN)'	15890.198	2
3	non-polymer	syn	'PROTOPORPHYRIN IX CONTAINING FE'	616.487	4
4	non-polymer	syn	'PHOSPHATE ION'	94.971	2
5	water	nat	water	18.015	221

mmCIFのキーバリュー形式は１行の表

キー・バリュー形式の場合も、「１行しかデータがない表」として読む。

```
#
_symmetry.entry_id                          2HHB
_symmetry.space_group_name_H-M              'P 1 21 1'
_symmetry.pdbx_full_space_group_name_H-M    ?
_symmetry.cell_setting                      ?
_symmetry.Int_Tables_number                 4
#
```

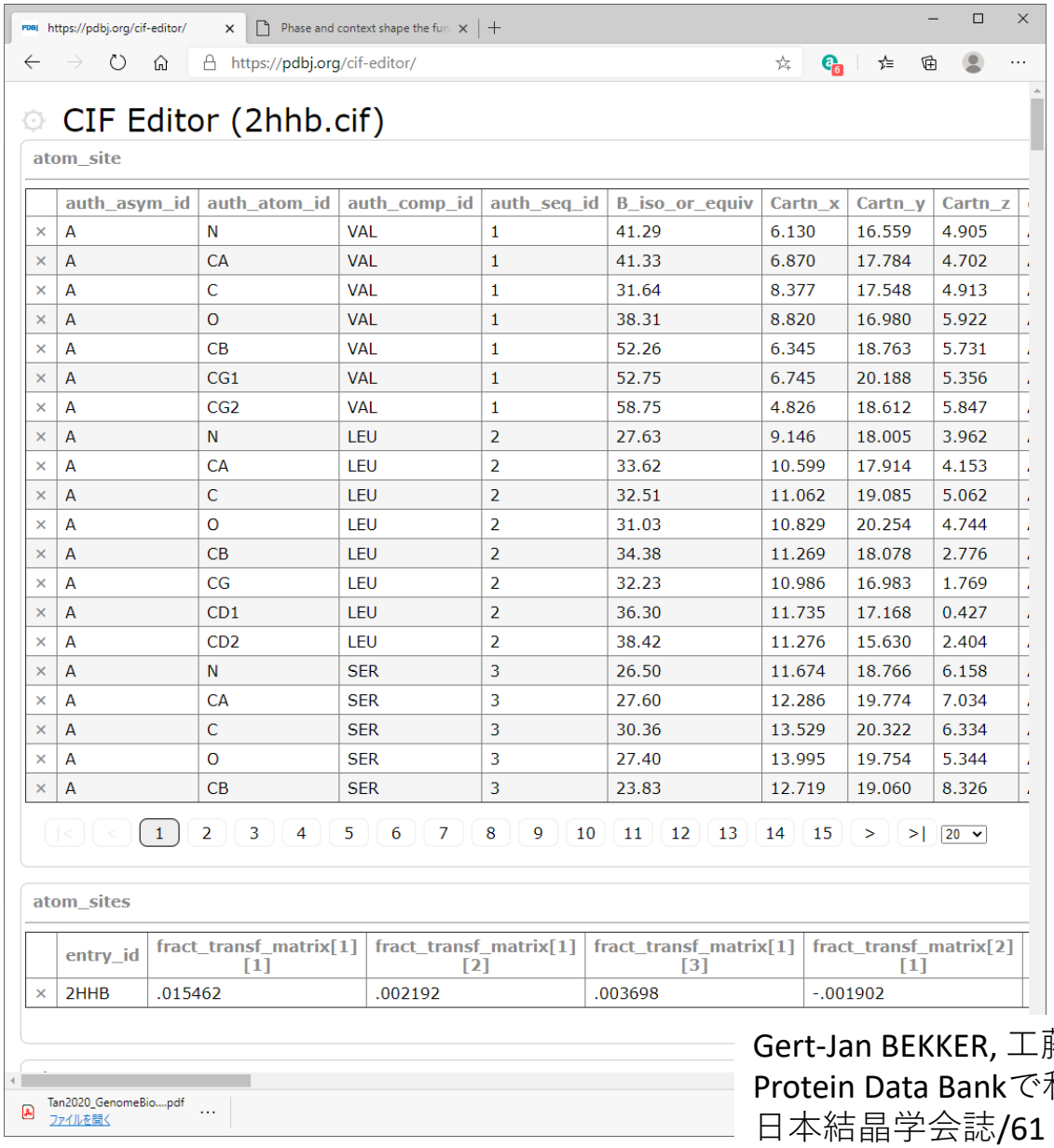
symmetry				
entry_id	space_group_name_H-M	pdbx_full_space_group_name_H-M	cell_setting	Int_Tables_number
2HHB	'P 1 21 1’	?	?	4

Pythonの場合：

D[表名][列名][行番号]

```
D[‘symmetry’][‘entity_id’][0]=‘2HHB’
D[‘entity’][‘space_group_name_H-M’][1]=‘P 1 21 1’
:
```

参考サイト:mmCIFエディタ



	auth_asym_id	auth_atom_id	auth_comp_id	auth_seq_id	B_iso_or_equiv	Cartn_x	Cartn_y	Cartn_z
×	A	N	VAL	1	41.29	6.130	16.559	4.905
×	A	CA	VAL	1	41.33	6.870	17.784	4.702
×	A	C	VAL	1	31.64	8.377	17.548	4.913
×	A	O	VAL	1	38.31	8.820	16.980	5.922
×	A	CB	VAL	1	52.26	6.345	18.763	5.731
×	A	CG1	VAL	1	52.75	6.745	20.188	5.356
×	A	CG2	VAL	1	58.75	4.826	18.612	5.847
×	A	N	LEU	2	27.63	9.146	18.005	3.962
×	A	CA	LEU	2	33.62	10.599	17.914	4.153
×	A	C	LEU	2	32.51	11.062	19.085	5.062
×	A	O	LEU	2	31.03	10.829	20.254	4.744
×	A	CB	LEU	2	34.38	11.269	18.078	2.776
×	A	CG	LEU	2	32.23	10.986	16.983	1.769
×	A	CD1	LEU	2	36.30	11.735	17.168	0.427
×	A	CD2	LEU	2	38.42	11.276	15.630	2.404
×	A	N	SER	3	26.50	11.674	18.766	6.158
×	A	CA	SER	3	27.60	12.286	19.774	7.034
×	A	C	SER	3	30.36	13.529	20.322	6.334
×	A	O	SER	3	27.40	13.995	19.754	5.344
×	A	CB	SER	3	23.83	12.719	19.060	8.326

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 > >| 20

	entry_id	fract_transf_matrix[1][1]	fract_transf_matrix[1][2]	fract_transf_matrix[1][3]	fract_transf_matrix[2][1]
×	2HHB	.015462	.002192	.003698	-.001902

https://pdbj.org/cif-editor/

・mmCIFファイルを読み込み、変更、修正ができる

・PDBx/mmCIF辞書に従ったmmCIFファイルの入力を支援

・mmCIFの内容をチェックするのにも便利

・キーバリュー形式のカテゴリも、１行の表として表現されている。

Gert-Jan BEKKER, 工藤 高裕, 池川 恭代, 山下 鈴子, 栗栖 源嗣
Protein Data Bankで利用するPDBx/mmCIF形式について
日本結晶学会誌/61 巻 (2019) 3 号/書誌 p. 159-160

atom_siteの表だけを読む方法

旧PDBフォーマットのATOM/HETATM行の情報は、atom_siteの表に全て書かれている。

```
#
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.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   . VAL A 1 1   ? 6.130   16.559  4.905   1.00 41.29 ? 1   VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? 6.870   17.784  4.702   1.00 41.33 ? 1   VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? 8.377   17.548  4.913   1.00 31.64 ? 1   VAL A C   1
ATOM   4    O  O   . VAL A 1 1   ? 8.820   16.980  5.922   1.00 38.31 ? 1   VAL A O   1
ATOM   5    C  CB  . VAL A 1 1   ? 6.345   18.763  5.731   1.00 52.26 ? 1   VAL A CB  1
ATOM   6    C  CG1 . VAL A 1 1   ? 6.745   20.188  5.356   1.00 52.75 ? 1   VAL A CG1 1
ATOM   7    C  CG2 . VAL A 1 1   ? 4.826   18.612  5.847   1.00 58.75 ? 1   VAL A CG2 1
ATOM   8    N  N   . LEU A 1 2   ? 9.146   18.005  3.962   1.00 27.63 ? 2   LEU A N   1
ATOM   9    C  CA  . LEU A 1 2   ? 10.599  17.914  4.153   1.00 33.62 ? 2   LEU A CA  1
ATOM   10   C  C   . LEU A 1 2   ? 11.062  19.085  5.062   1.00 32.51 ? 2   LEU A C   1
:
HETATM 4385 C  CHA . HEM E 3 .   ? 8.634   7.898   -18.334 1.00 17.24 ? 142 HEM A CHA 1
HETATM 4386 C  CHB . HEM E 3 .   ? 10.386  9.965   -14.276 1.00 23.17 ? 142 HEM A CHB 1
HETATM 4387 C  CHC . HEM E 3 .   ? 8.337   6.403   -11.611 1.00 11.04 ? 142 HEM A CHC 1
:
#
```

データベース管理者が入力する

原子座標

登録者が入力する

旧PDBフォーマットの
ATOM/HETATM行に相当する列

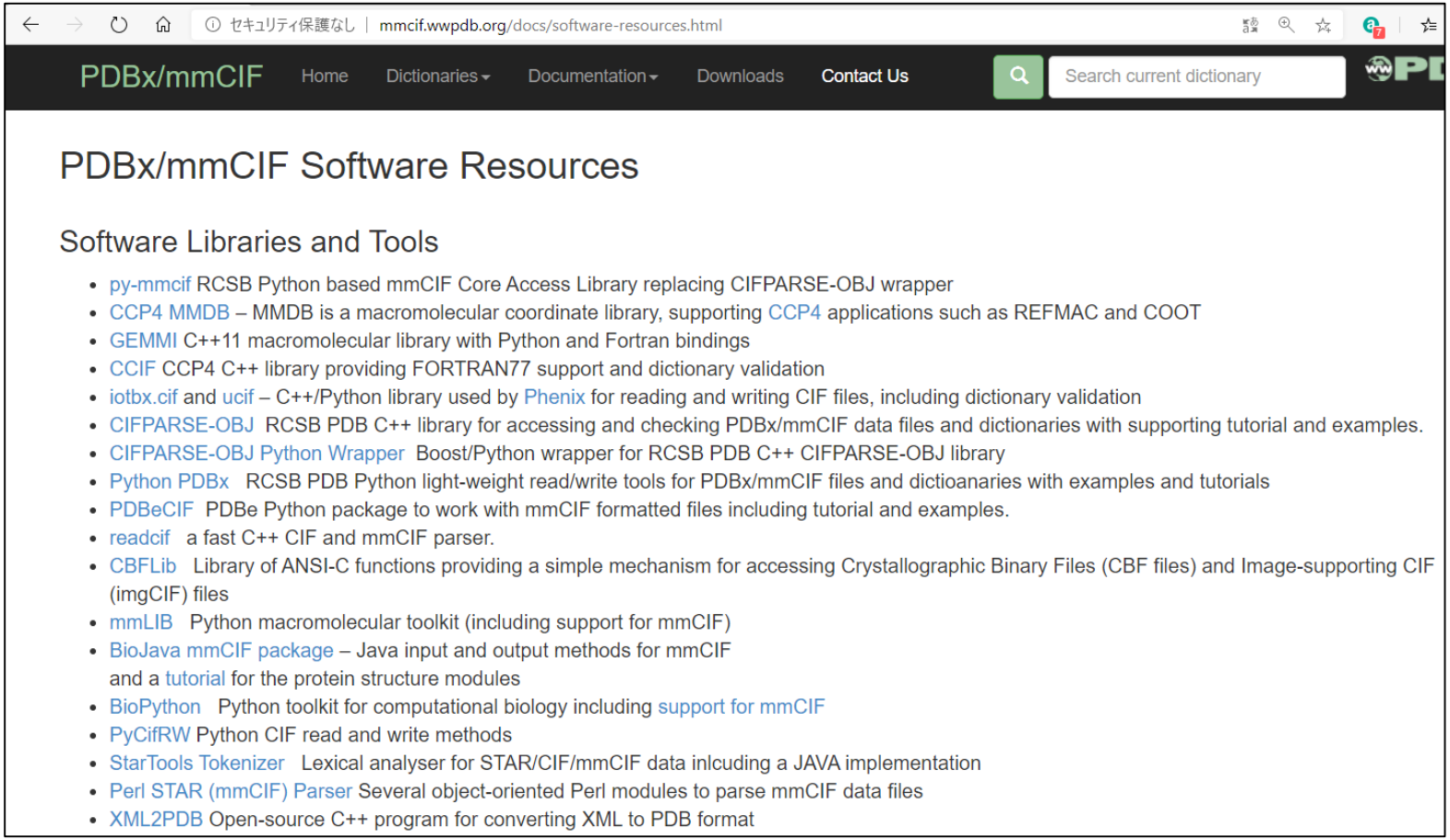
原子番号　原子名　残基名　残基番号　X座標　原子番号　Y座標　Z座標　占有率　温度因子　残基番号　asym_id(鎖識別子)

元素名　asym_id　残基名　原子名

ライブラリを使用する方法

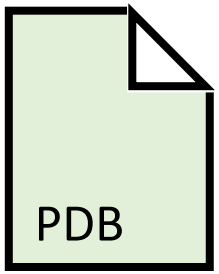
英語版： http://mmcif.wwpdb.org/docs/software-resources.html

日本語版：https://mmcif.pdbj.org/docs/software-resources.html



ファイルフォーマットの詳細情報

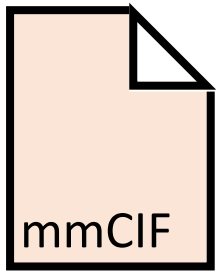
旧PDBフォーマットについて



http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html

mmCIFフォーマットについて (英語版)

http://mmcif.wwpdb.org/



mmCIFフォーマットについて (日本語版)

https://mmcif.pdbj.org/

※ PDB Exchange Dictionary (PDBx/mmCIF)に詳細な定義が書かれている

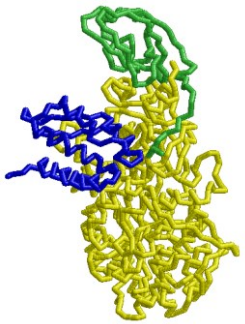
PDBj 20 Year Anniversary

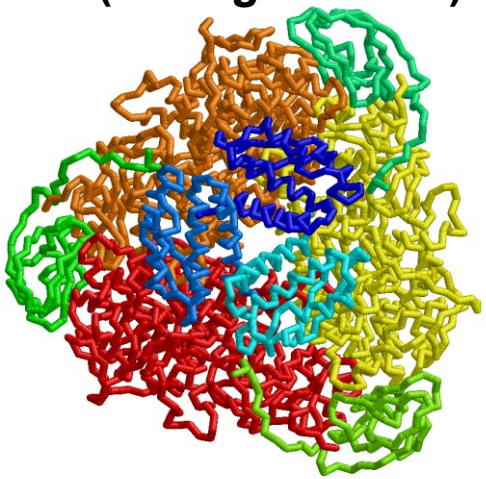
ご清聴ありがとうございました

非対称単位と 生物学的単位

(*Asymmetric Unit*) **(*Biological Unit*)**

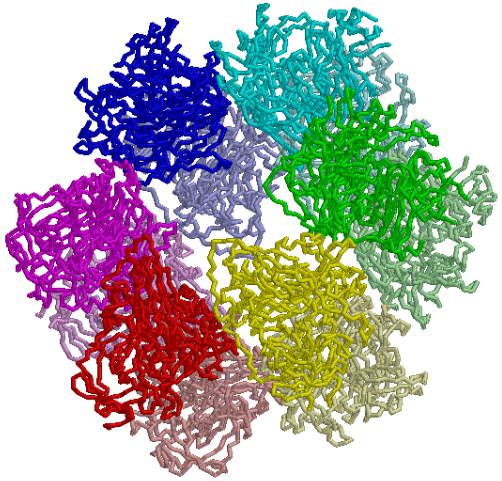
PDBcode:1a5k



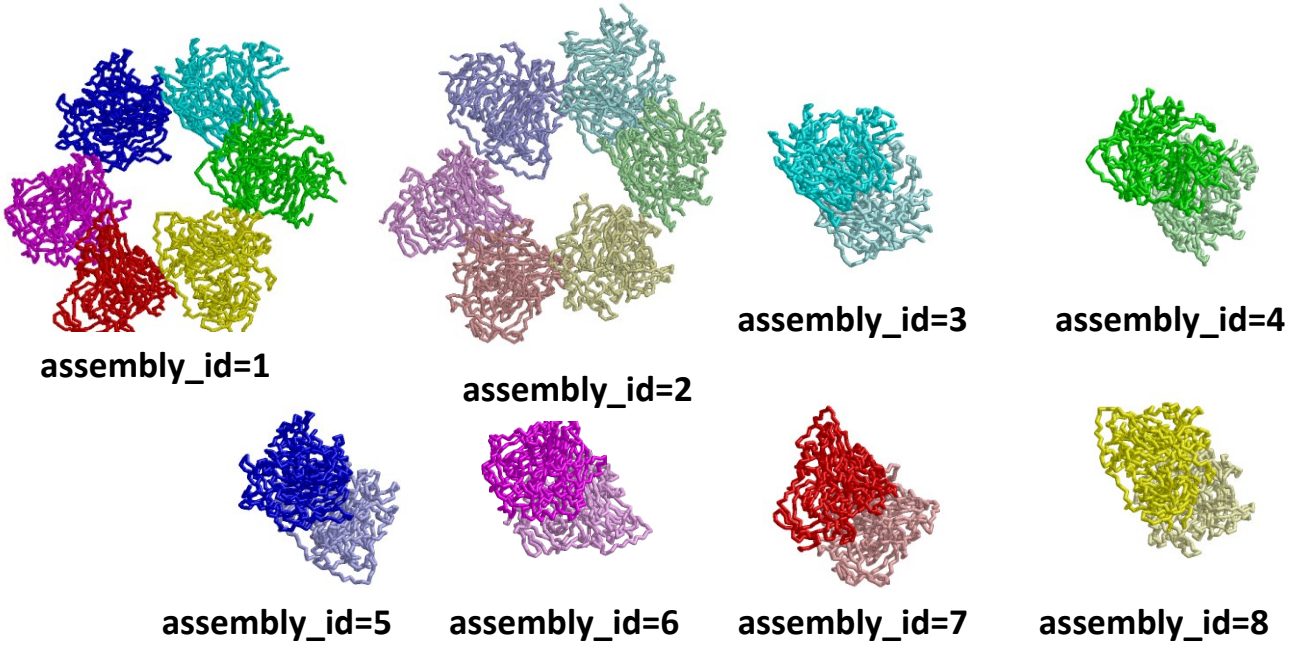


Asymmetric Unit ***Biological Unit***

PDBcode:3gyr



Asymmetric Unit



Biological Unit

Biological Unit の記載例

label_asym_id=A,B,C:蛋白質
label_asym_id=D,E,F:水分子

mmCIF フォーマット(*1a5k*)

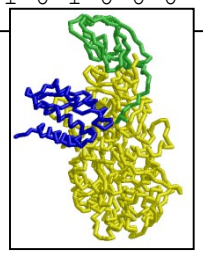
```
_pdbx_struct_assembly.id                    1
_pdbx_struct_assembly.details               author_and_software_defined_assembly
_pdbx_struct_assembly.method_details        PISA,PQS
_pdbx_struct_assembly.oligomeric_details    nonameric
_pdbx_struct_assembly.oligomeric_count      9
#
_pdbx_struct_assembly_gen.assembly_id        1
_pdbx_struct_assembly_gen.oper_expression    1,2,3
_pdbx_struct_assembly_gen.asym_id_list       A,D,B,E,C,F
#
loop_
_pdbx_struct_oper_list.id
_pdbx_struct_oper_list.type
_pdbx_struct_oper_list.name
_pdbx_struct_oper_list.symmetry_operation
_pdbx_struct_oper_list.matrix[1][1]
_pdbx_struct_oper_list.matrix[1][2]
_pdbx_struct_oper_list.matrix[1][3]
_pdbx_struct_oper_list.vector[1]
_pdbx_struct_oper_list.matrix[2][1]
_pdbx_struct_oper_list.matrix[2][2]
_pdbx_struct_oper_list.matrix[2][3]
_pdbx_struct_oper_list.vector[2]
_pdbx_struct_oper_list.matrix[3][1]
_pdbx_struct_oper_list.matrix[3][2]
_pdbx_struct_oper_list.matrix[3][3]
_pdbx_struct_oper_list.vector[3]
1 'identity operation'          1_555 x,y,z 1 0 0 0 0 1 0 0 0 0 1 0
2 'crystal symmetry operation' 5_555 z,x,y 0 0 1 0 1 0 0 0 0 1 0 0
3 'crystal symmetry operation' 9_555 y,z,x 0 1 0 0 0 0 1 0 1 0 0 0
```

oper_expression : 座標変換操作名

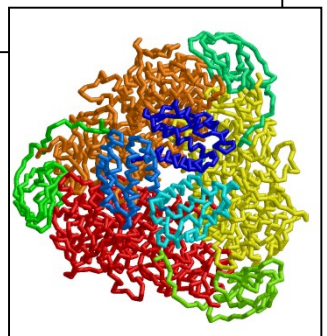
従来のPDB フォーマット (*1a5k*)

```
REMARK 300 BIOMOLECULE: 1
REMARK 300 SEE REMARK 350 FOR THE AUTHOR PROVIDED AND/OR PROGRAM
REMARK 300 GENERATED ASSEMBLY INFORMATION FOR THE STRUCTURE IN
REMARK 300 THIS ENTRY. THE REMARK MAY ALSO PROVIDE INFORMATION ON
REMARK 300 BURIED SURFACE AREA.
REMARK 350
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW.  BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN.
REMARK 350
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: NONAMERIC
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT1   2  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT2   2  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT3   2  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT1   3  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT2   3  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT3   3  1.000000  0.000000  0.000000        0.00000
```

座標変換操作名(**oper_expression**)と
その変換を施す分子名(**asym_id**)の組合せで
biological unitが記述されている。



Asymmetric unit

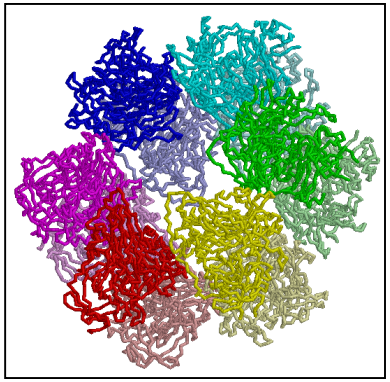


Biological unit
(assembly_id=1)

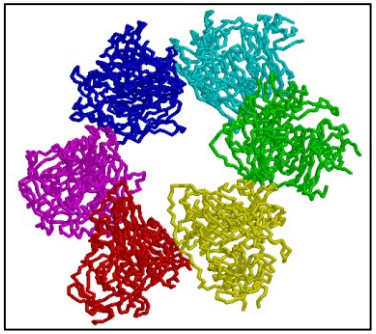
Biological Unitの記載例

複数のBiological unitがある場合は、それぞれ別の **assembly_id** が振られている

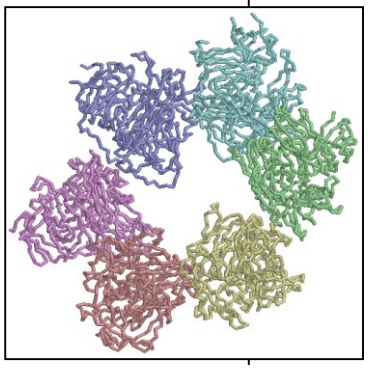
```
loop_
_pdbx_struct_assembly.id
_pdbx_struct_assembly.details
_pdbx_struct_assembly.method_details
_pdbx_struct_assembly.oligomeric_details
_pdbx_struct_assembly.oligomeric_count
1 author_defined_assembly   ?       hexameric 6
2 author_defined_assembly   ?       hexameric 6
3 software_defined_assembly PISA,PQS dimeric   2
4 software_defined_assembly PISA,PQS dimeric   2
5 software_defined_assembly PISA,PQS dimeric   2
6 software_defined_assembly PISA,PQS dimeric   2
7 software_defined_assembly PISA,PQS dimeric   2
8 software_defined_assembly PISA,PQS dimeric   2
#
loop_
_pdbx_struct_assembly_gen.assembly_id
_pdbx_struct_assembly_gen.oper_expression
_pdbx_struct_assembly_gen.asym_id_list
1 1 A,M,N,O,P,Q,UB,B,R,S,T,U,V,VB,C,W,X,Y,Z,AA,WB,D,BA,CA,DA,EA,FA,XB,E,GA,HA,IA,JA,KA,YB,F,LA,MA,NA,OA,PA,ZB
2 1 G,QA,RA,SA,TA,UA,AC,H,VA,WA,XA,YA,ZA,BC,I,AB,BB,CB,DB,EB,CC,J,FB,GB,HB,IB,JB,DC,K,KB,LB,MB,NB,OB,EC,L,PB,QB,RB,SB,TB,FC
3 1 F,LA,MA,NA,OA,PA,ZB,J,FB,GB,HB,IB,JB,DC
4 1 E,GA,HA,IA,JA,KA,YB,K,KB,LB,MB,NB,OB,EC
5 1 A,M,N,O,P,Q,UB,I,AB,BB,CB,DB,EB,CC
6 1 B,R,S,T,U,V,VB,H,VA,WA,XA,YA,ZA,BC
7 1 C,W,X,Y,Z,AA,WB,G,QA,RA,SA,TA,UA,AC
8 1 D,BA,CA,DA,EA,FA,XB,L,PB,QB,RB,SB,TB,FC
#
_pdbx_struct_oper_list.id                    1
_pdbx_struct_oper_list.type                  'identity operation'
_pdbx_struct_oper_list.name                  1_555
_pdbx_struct_oper_list.symmetry_operation    x,y,z
```



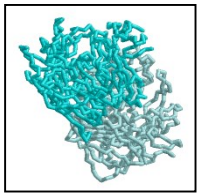
Asymmetric unit



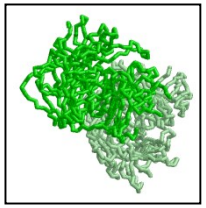
assembly_id=1



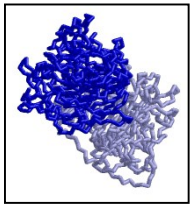
assembly_id=2



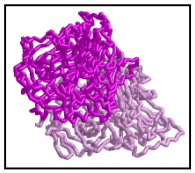
assembly_id=3



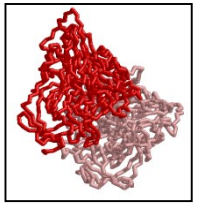
assembly_id=4



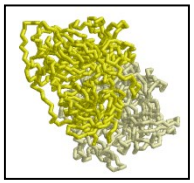
assembly_id=5



assembly_id=6



assembly_id=7



assembly_id=8

PDBcode:3gyr