

PDBj Mine の使い方

PDBj講習会 in 長浜

金城玲
大阪大学蛋白質研究所
2011年1月17日

1

この講義の内容

- PDBデータの読み方
 - PDB format
 - mmCIF format
 - PDBML format
- 演習
 - 実際にどのようなデータがどのように書かれているかを調べる。

2

基本的な検索法

- エントリーの取得
- キーワード検索
- Advanced search (細かな条件を指定した検索)

<http://www.pdbj.org/>

を開いてください。

3

PDBjのウェブインターフェイス



- データの登録(今回は扱いません)。
- データの検索(エントリー／キーワード)
- その他のサービス(ページ左側)。
- ページの右上に現在のエントリー数が表示されている。(クリックすると「今週追加／更新されたエントリー」のリストが表示される。

4

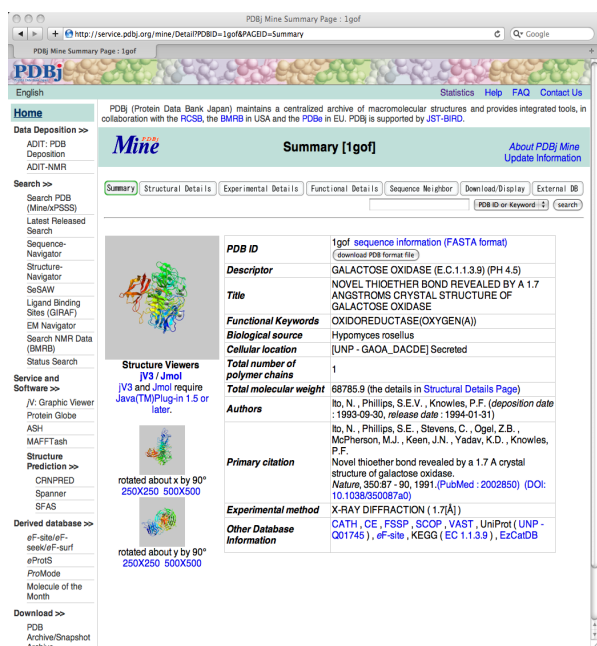
エントリーの取得



- 「PDB ID」がチェックされていることを確認する。
- 箱に「1gof」と入力する（半角英数字）。
- 「Go」ボタンをクリック。
- 「1gof」はあるエントリーのPDB ID。

5

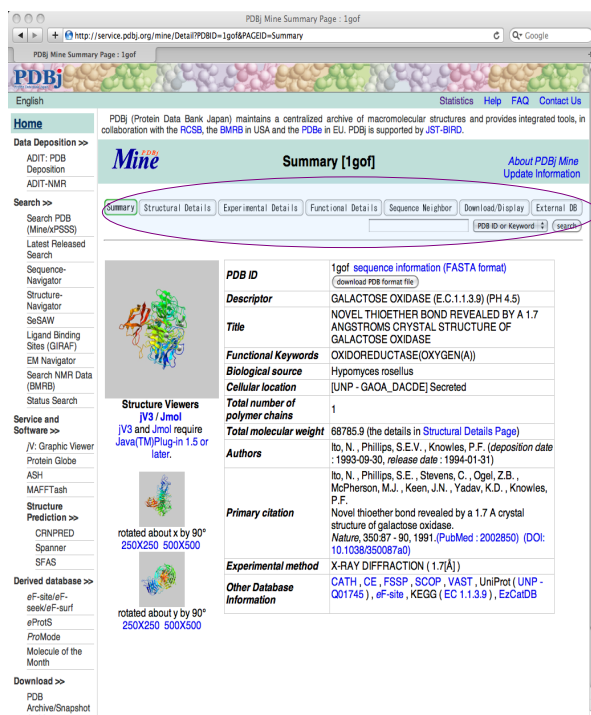
エントリーのサマリー(要約)ページ



- エントリー「1gof」のサマリーページが表示されます。
- サマリーとはこの蛋白質に関する記述、引用文献、実験手法、他のデータベースの該当するエントリーへのリンク、などをまとめたものの。
- 図もついています。

6

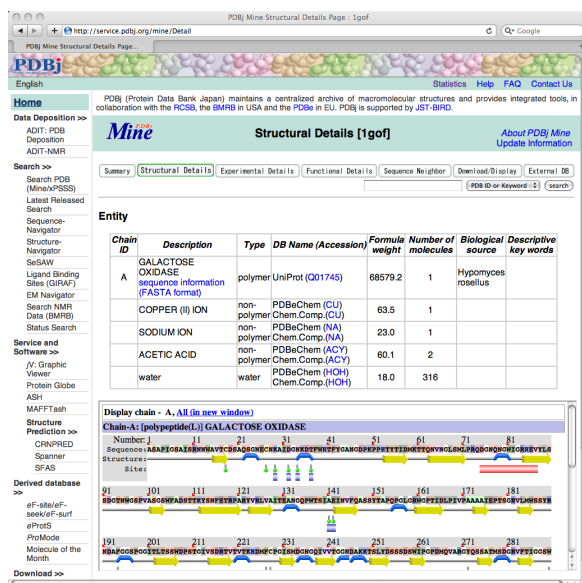
エントリーのその他の情報



- サマリーページ上部にその他の情報のページへのリンクがあります。
- Structural Details
- Experimental Details
- Functional Details
- Sequence Neighbor
- Download/Display
- Links
- これらを順に説明します。

7

Structural Details



- 解かれている構造の詳細情報です。
- 「どのようなモノがふくまれるか」が主な内容です。
- 1gofの場合はポリペプチド鎖が1つと銅イオン、ナトリウムイオン、酢酸、水、が含まれていることがわかります。

8

Experimental Details

Experimental Details (X-Ray) [1go1]

Summary | Structural Details | **Experimental Details** | Functional Details | Sequence Neighbor | Download/Display | External DB

Refinement Statistics

Cell axes [Å]	98.000	89.400	86.700
Cell angles (degree)	90.00	117.80	90.00
Spacegroup	C 1 2 1		
Resolution limits [Å]	10.0 ¹ - 1.7 ¹		
R-factor	0.177		
R-free			
RMSD bond length [Å]	0.018		
RMSD bond angle [Å]			

Values in parentheses refer to the highest resolution shell.

Data Collection Statistics

Resolution limits [Å]	10.0 ¹ - 1.7 ¹ (1.8 ² - 1.7)
Number of reflections	57092 ² (6908) ²
Number of measurements	117980 ² (11365) ²
Rmerge I_obs	0.039 ²
Completeness [%]	79.1 ² (82.6) ²
Redundancy	

Values in parentheses refer to the highest resolution shell.

Crystallization Conditions

method	pH	temperature
unknown ¹		

Crystallization Reagents

- 構造が解かれた実験の条件に関する詳細です。
- 結晶構造の場合は結晶の各種パラメータや結晶化条件、溶液 NMR構造の場合は分光器やスペクトルの種類などが記載されています。

9

Functional Details

Functional Details [1go1]

Summary | Structural Details | Experimental Details | **Functional Details** | Sequence Neighbor | Download/Display | External DB

about Functional Details Page

Functional Information from GO Data

Chain	GOid	namespace	contents
A	0005507	molecular_function	copper ion binding
A	0005576	cellular_component	extracellular region
A	0007155	biological_process	cell adhesion
A	0045480	molecular_function	galactose oxidase activity
A	0055114	biological_process	oxidation reduction

(More GO Information ...)

Functional Information from PDB Data

site_id	type	Number of Residues	Details
CU	SITE	5	COPPER BINDING SITE
NA	SITE	6	SODIUM BINDING SITE
AC1	SITE	7	BINDING SITE FOR RESIDUE CU A 700
AC2	SITE	6	BINDING SITE FOR RESIDUE NA A 702
AC3	SITE	7	BINDING SITE FOR RESIDUE ACY A 701
AC4	SITE	9	BINDING SITE FOR RESIDUE ACY A 703

(More PDB Information ...)

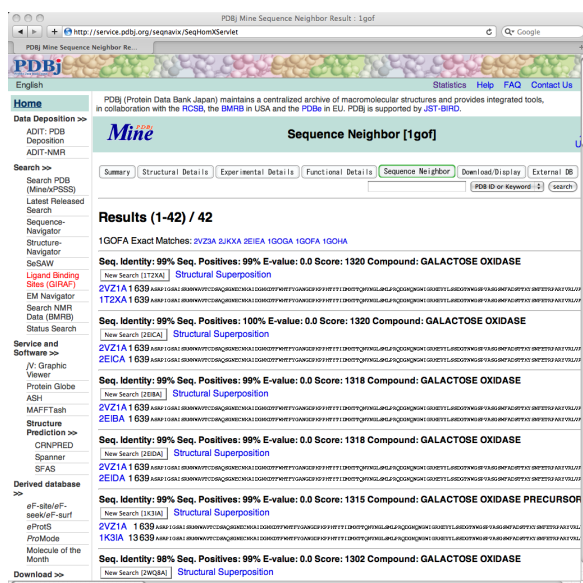
Functional Information from PDB atom coordinates for the "HETATM" binding sites

site_id	type	Number of Residues	Details
ACY_1go1_A_703	binding site	11	ACETIC ACID binding site
ACY_1go1_A_701	binding site	11	ACETIC ACID binding site

- 蛋白質の機能に関する情報が記載されています。
- 多くは、PDBjが独自に収集したデータで元のPDBのデータにはない情報も含まれます。
- Gene ontology
- Functional site (残基)
- リガンド結合部位
 - などなど

10

Sequence Neighbor

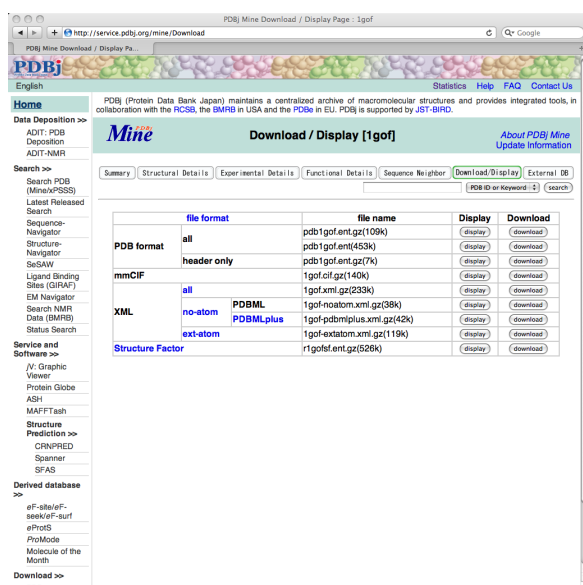


- PDBに対する配列類似性検索 (BLAST) の結果を表示します。

- 注意: 複数のポリペプチド鎖が含まれるエントリーの場合は最初のポリペプチド鎖に対して検索されます。

11

Download/Display



- 元データを表示またはダウンロードできます。
- 「生」のPDB format, mmCIF, PDBMLフォーマットの他に、原子座標を除いたバージョンも用意されています。
- アノテーションのみがみたいときに便利
- 結晶構造の場合は構造因子も用意されていることもあります。
- 後ほどじっくり眺めます。

12

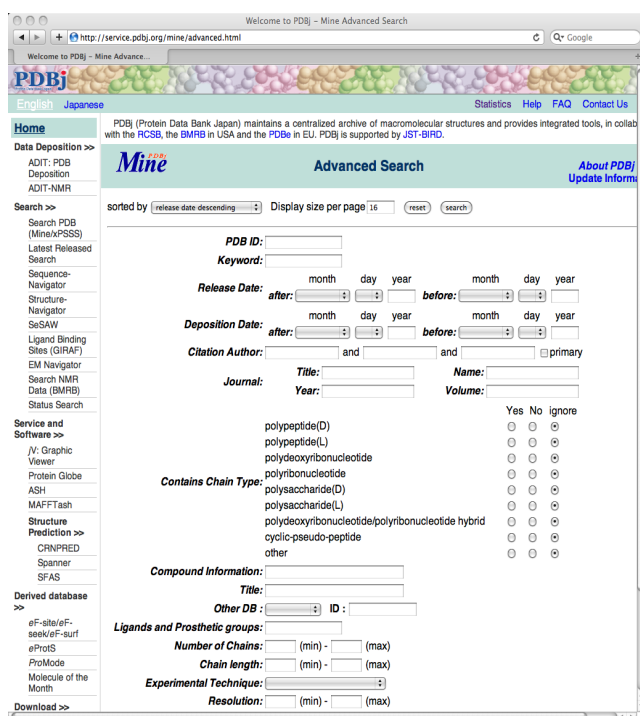
Advanced Search



- 再びトップページに戻り、「Advanced Search」というリンクをクリック。
- より詳細な条件を指定した検索ができます。

13

Advanced Searchを使う



- 以下の条件などで検索できます
 - 登録/リリース年月日
 - 登録者(論文著者)名
 - 論文のタイトル、雑誌名
 - 含まれるポリマーの種類(ポリペプチド、DNA, RNA, etc.)
 - 解像度
 - リガンド分子名
 - (後でみっちり練習します)

14

データバンクとウェブサイト

- 「ウェブサイト(例: www.pdbj.org)がデータバンク(例: PDB)」ではない。
- データバンクとは1次データが集積されている場所(組織)のこと(PDBやINSO[核酸配列])。
- PDBの基本はデータバンク(≠データベース)。
- PDBj内に構築されたデータベースはウェブサイト経由でしか使えない。
- ウェブサイトで表示されるものがデータバンクに含まれるデータのすべてではない。
- 詳しい情報はエントリーのファイルを直接みる必要がある。

15

エントリーとエントリーファイル

- 「エントリー」とは一つの実験で決まったひとまとまりの立体構造の情報のこと。
- 「PDB ID」は各エントリーに固有の識別子で英数字4文字からなる。例: 1a00, 1gof (大文字・小文字の区別はない)
- 各エントリーのファイルはwww.pdbj.org などからダウンロードできる。
- エントリーファイルにはそのエントリーに関する(ほとんど)すべての情報が含まれる。(ただし、各種 convention の定義などは除く。)

16

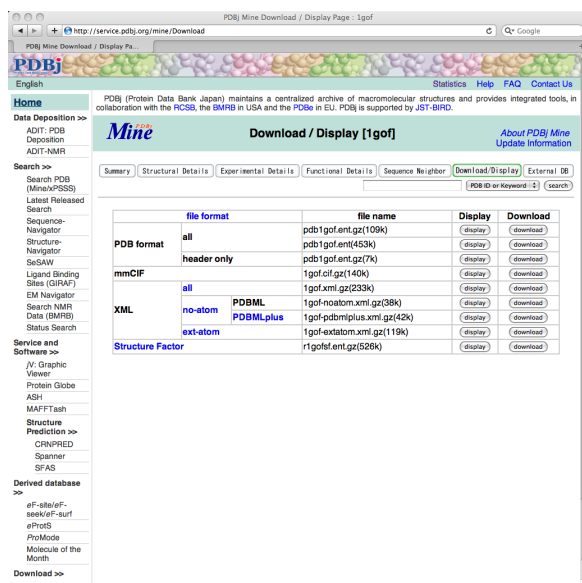
PDBのファイル形式

各PDBエントリーは一つのファイルに収められている。

- PDB フォーマット
 - 伝統的なフラットファイル形式
 - 情報に欠落があるので厳密な解析には向かない(簡便フォーマット)。
- mmCIF
 - 国際結晶学連合(IUCr)が定めたCIF(Crystallographic Information File)を拡張したもの
 - Hall, S.R., et al. (1991), *Acta Cryst. A*, **47**:655-685.
- PDBML
 - mmCIFに基づくXMLフォーマット
 - Westbrook, J. et al. (2005), *Bioinformatics* **21**:988-992

17

「Download/Display」をもう一度開く



- 「Display」ボタンを押すとブラウザでファイルの中身が表示されます。
- ただし、XML(no-atom)だけは例外。

18

PDBフォーマット

```

HEADER      OXIDOREDUCTASE(OXYGEN(A))                      30-SEP-93      1GOF
TITLE       NOVEL THIOETHER BOND REVEALED BY A 1.7 ANGSTROMS CRYSTAL
TITLE       2 STRUCTURE OF 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      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.V. PHILLIPS, C. STEVENS, Z.B. OGEL,
JRNL        AUTH 2 M.J. MCPHERSON, J.N. KEEN, K.D. YADAV, P.F. KNOWLES
JRNL        TITL NOVEL THIOETHER BOND REVEALED BY A 1.7 A CRYSTAL
JRNL        TITL 2 STRUCTURE 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,
REMARK      1 TITL 2 GALACTOSE 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,
REMARK      1 AUTH 2 J.M. KEEN, P.F. KNOWLES
REMARK      1 TITL GALACTOSE OXIDASE OF DACTYLIUM DENDROIDES: GENE
REMARK      1 TITL 2 CLONING AND 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.

```

19

PDBフォーマット(つづき)

```

ATOM       1 N ALA A 1 38.840 0.236 1.012 1.00 34.65 N
ATOM       2 CA ALA A 1 38.356 -0.999 0.357 1.00 42.26 C
ATOM       3 C ALA A 1 37.098 -1.547 1.056 1.00 41.25 C
ATOM       4 O ALA A 1 36.619 -0.946 2.028 1.00 29.44 O
ATOM       5 CB ALA A 1 39.398 -2.114 0.379 1.00 40.70 C
ATOM       6 N SER A 2 36.610 -2.666 0.495 1.00 32.67 N
ATOM       7 CA SER A 2 35.411 -3.244 1.202 1.00 34.90 C
ATOM       8 C SER A 2 35.683 -4.740 1.081 1.00 38.30 C
ATOM       9 O SER A 2 36.827 -5.147 0.747 1.00 28.59 O
ATOM      10 CB SER A 2 34.063 -2.660 0.823 1.00 24.49 C
ATOM      11 OG SER A 2 33.031 -3.308 1.686 1.00 20.37 O
ATOM      12 N ALA A 3 34.660 -5.537 1.334 1.00 35.91 N
ATOM      13 CA ALA A 3 34.815 -6.995 1.246 1.00 33.38 C
ATOM      14 C ALA A 3 33.416 -7.594 1.259 1.00 21.71 C
ATOM      15 O ALA A 3 32.529 -6.833 1.679 1.00 30.82 O
ATOM      16 CB ALA A 3 35.687 -7.414 2.433 1.00 25.44 C
ATOM      17 N PRO A 4 33.335 -8.786 0.733 1.00 36.18 N
ATOM      18 CA PRO A 4 32.068 -9.552 0.674 1.00 40.82 C
ATOM      19 C PRO A 4 32.067 -10.418 1.934 1.00 37.76 C
ATOM      20 O PRO A 4 33.145 -10.698 2.466 1.00 42.84 O
ATOM      21 CB PRO A 4 32.222 -10.479 -0.536 1.00 40.12 C
ATOM      22 CG PRO A 4 33.729 -10.691 -0.634 1.00 48.00 C
ATOM      23 CD PRO A 4 34.452 -9.579 0.148 1.00 34.36 C

```

20

PDBフォーマット超入門

- 行指向(1行80コラム)(cf. FORTRAN77)
- ヘッダ(行頭6コラム)とアノテーションの組。
 - 長いアノテーションは次の行に継続。(7~10コラム目に継続を示す数字がある)
- 代表的なヘッダは: HEADER, TITLE, COMPND, SOURCE, EXPDTA, AUTHOR, JRNL, REMARK (番号つき), ATOM, HETATM
- 詳しくは <http://www.wwpdb.org/docs.html>

21

mmCIF

```

data_1GOF
#
_entry.id      1GOF
#
_audit_conform.dict_name      mmcif_pdbx.dic
_audit_conform.dict_version   1.0670
_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
#
loop_
_database_PDB_rev_record.rev_num
_database_PDB_rev_record.type
_database_PDB_rev_record.details
2 JRNL ?
3 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.SG_entry         .
#
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Ito, N.' 1
'Phillips, S.E.V.' 2
'Knowles, P.F.' 3
#

```

22

```

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

```

mmCIF (つづき)

```

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
ATOM 12 N N . ALA A 1 3 ? 34.660 -5.537 1.334 1.00 35.91 ? ? ? ? ? 3 ALA A N 1
ATOM 13 C CA . ALA A 1 3 ? 34.815 -6.995 1.246 1.00 33.38 ? ? ? ? ? 3 ALA A CA 1
ATOM 14 C C . ALA A 1 3 ? 33.416 -7.594 1.259 1.00 21.71 ? ? ? ? ? 3 ALA A C 1
ATOM 15 O O . ALA A 1 3 ? 32.529 -6.833 1.679 1.00 30.82 ? ? ? ? ? 3 ALA A O 1
ATOM 16 C CB . ALA A 1 3 ? 35.687 -7.414 2.433 1.00 25.44 ? ? ? ? ? 3 ALA A CB 1
ATOM 17 N N . PRO A 1 4 ? 33.335 -8.786 0.733 1.00 36.18 ? ? ? ? ? 4 PRO A N 1
ATOM 18 C CA . PRO A 1 4 ? 32.068 -9.552 0.674 1.00 40.82 ? ? ? ? ? 4 PRO A CA 1
ATOM 19 C C . PRO A 1 4 ? 32.067 -10.418 1.934 1.00 37.76 ? ? ? ? ? 4 PRO A C 1
ATOM 20 O O . PRO A 1 4 ? 33.145 -10.698 2.466 1.00 42.84 ? ? ? ? ? 4 PRO A O 1
ATOM 21 C CB . PRO A 1 4 ? 32.222 -10.479 -0.536 1.00 40.12 ? ? ? ? ? 4 PRO A CB 1
ATOM 22 C CG . PRO A 1 4 ? 33.729 -10.691 -0.634 1.00 48.00 ? ? ? ? ? 4 PRO A CG 1
ATOM 23 C CD . PRO A 1 4 ? 34.452 -9.579 0.148 1.00 34.36 ? ? ? ? ? 4 PRO A CD 1

```

23

mmCIF超入門

- 文脈自由文法(STAR[Self-defining Text Archive and Retrieval]形式)
- タグとアノテーションの組。
- タグは「_」で始まる。
- タグとアノテーションの組が複数ある場合には「loop_」構文で繰り返しを指定する。
- タグの意味はmmCIF dictionary(STAR形式)で定義されている。
- 色々な種類のデータは「カテゴリー」によって分類・整理されている。
 - カテゴリーの例: entity, struct, citation, citation_author, atom_site, chem_comp, cell, ...
- 詳しくは <http://mmcif.pdb.org/>

24

PDBML

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock datablockName="1GOF-noatom"
  xmlns:PDBx="http://pdbml.pdb.org/schema/pdbx-v32.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://pdbml.pdb.org/schema/pdbx-v32.xsd pdbx-v32.xsd">
  <PDBx:atom_sitesCategory>
    <PDBx:atom_sites entry_id="1GOF">
      <PDBx:Cartn_transform_axes xsi:nil="true" />
      <PDBx:fract_transf_matrix11>0.011535</PDBx:fract_transf_matrix11>
      <PDBx:fract_transf_matrix12>0.000000</PDBx:fract_transf_matrix12>
      <PDBx:fract_transf_matrix13>0.000000</PDBx:fract_transf_matrix13>
      <PDBx:fract_transf_matrix21>0.000000</PDBx:fract_transf_matrix21>
      <PDBx:fract_transf_matrix22>0.011186</PDBx:fract_transf_matrix22>
      <PDBx:fract_transf_matrix23>0.000000</PDBx:fract_transf_matrix23>
      <PDBx:fract_transf_matrix31>0.006081</PDBx:fract_transf_matrix31>
      <PDBx:fract_transf_matrix32>0.000000</PDBx:fract_transf_matrix32>
      <PDBx:fract_transf_matrix33>0.011534</PDBx:fract_transf_matrix33>
      <PDBx:fract_transf_vector1>0.00000</PDBx:fract_transf_vector1>
      <PDBx:fract_transf_vector2>0.00000</PDBx:fract_transf_vector2>
      <PDBx:fract_transf_vector3>0.00000</PDBx:fract_transf_vector3>
    </PDBx:atom_sites>
  </PDBx:atom_sitesCategory>
  <PDBx:atom_sites_footnoteCategory>
    <PDBx:atom_sites_footnote id="1">
      <PDBx:text>CIS PROLINE - PRO 52</PDBx:text>
    </PDBx:atom_sites_footnote>
    <PDBx:atom_sites_footnote id="2">
      <PDBx:text>CIS PROLINE - PRO 163</PDBx:text>
    </PDBx:atom_sites_footnote>
    <PDBx:atom_sites_footnote id="3">
      <PDBx:text>CIS PROLINE - PRO 350</PDBx:text>
    </PDBx:atom_sites_footnote>
  </PDBx:atom_sites_footnoteCategory>
```

25

PDBML (つづき)

```
<PDBx:atom_siteCategory>
  <PDBx:atom_site id="1">
    <PDBx:B_iso_or_equiv>34.65</PDBx:B_iso_or_equiv>
    <PDBx:B_iso_or_equiv_esd xsi:nil="true" />
    <PDBx:Cartn_x>38.840</PDBx:Cartn_x>
    <PDBx:Cartn_x_esd xsi:nil="true" />
    <PDBx:Cartn_y>0.236</PDBx:Cartn_y>
    <PDBx:Cartn_y_esd xsi:nil="true" />
    <PDBx:Cartn_z>1.012</PDBx:Cartn_z>
    <PDBx:Cartn_z_esd xsi:nil="true" />
    <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></PDBx:label_alt_id>
    <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:occupancy_esd xsi:nil="true" />
    <PDBx:pdxb_PDB_ins_code xsi:nil="true" />
    <PDBx:pdxb_PDB_model_num>1</PDBx:pdxb_PDB_model_num>
    <PDBx:pdxb_formal_charge xsi:nil="true" />
    <PDBx:type_symbol>N</PDBx:type_symbol>
  </PDBx:atom_site>

  <PDBx:atom_site id="2">
    <PDBx:B_iso_or_equiv>42.26</PDBx:B_iso_or_equiv>
    <PDBx:B_iso_or_equiv_esd xsi:nil="true" />
    <PDBx:Cartn_x>38.356</PDBx:Cartn_x>
    <PDBx:Cartn_x_esd xsi:nil="true" />
    <PDBx:Cartn_y>-0.999</PDBx:Cartn_y>
    <PDBx:Cartn_y_esd xsi:nil="true" />
    <PDBx:Cartn_z>0.357</PDBx:Cartn_z>
    <PDBx:Cartn_z_esd xsi:nil="true" />
    <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
    <PDBx:auth_atom_id>CA</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></PDBx:label_alt_id>
    <PDBx:label_asym_id>A</PDBx:label_asym_id>
    <PDBx:label_atom_id>CA</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:occupancy_esd xsi:nil="true" />
    <PDBx:pdxb_PDB_ins_code xsi:nil="true" />
    <PDBx:pdxb_PDB_model_num>1</PDBx:pdxb_PDB_model_num>
    <PDBx:pdxb_formal_charge xsi:nil="true" />
    <PDBx:type_symbol>C</PDBx:type_symbol>
  </PDBx:atom_site>
```

26

PDBML超入門

- 文脈自由文法(XML)
- mmCIFのカテゴリーに対応したXMLのタグ。
- XML要素の型はPDBMLのXML Schemaで定義されている。(タグの意味はmmCIFから継承されている。)

27

まとめ:フォーマットの比較

	形式	データの完全性	人が読みやすい	プログラムで処理しやすい	標準化
PDB	フラットファイル	×	◎	×	wwPDB
mmCIF	STAR	○	○	○	IUCr, wwPDB
PDBML	XML	○	×	◎	W3C, wwPDB

28

アンケートから(1)

- 「PDB fileを見ると、1つのchainに複数の低分子リガンドが入っているものがあるが、全てべつのchainとして登録してもらえるとありがたいです。」

→PDB ファイルではなく、mmCIFまたはPDBMLを参照すればよい。

- PDBファイルの“chain ID”は auth_asym_id
- mmCIF/PDBMLにはそれとは別に label_asym_idがある。これは異なる分子には異なるIDが割り当てられる。

29

分子の「chain ID」

1A00を例にとると……

PDB file

```
ATOM      1  N   VAL  A   1      101.601  38.534  -1.962  1.00  53.29      N
HETATM 4387  FE   HEM  A  142     93.518  58.468  -3.212  1.00  18.42      FE
```

mmCIF

```
ATOM      1      N  N   . VAL  A  1  1   ? 101.601 38.534 -1.962  1.00 53.29  ? ? ? ? ? 1   VAL  A  N   1
HETATM 4383  FE  FE   . HEM  E  3  .   ? 93.518 58.468 -3.212  1.00 18.42  ? ? ? ? ? 142 HEM  A  FE   1
```

PDBML

```
<PDBx:atom_site id="1">
```

```
...
<PDBx:auth_asym_id>A</PDBx:auth_asym_id>
<PDBx:auth_atom_id>N</PDBx:auth_atom_id>
<PDBx:auth_comp_id>VAL</PDBx:auth_comp_id>
<PDBx:auth_seq_id>1</PDBx:auth_seq_id>
<PDBx:group_PDB>ATOM</PDBx:group_PDB>
<PDBx:label_asym_id>A</PDBx:label_asym_id>
<PDBx:label_atom_id>N</PDBx:label_atom_id>
<PDBx:label_comp_id>VAL</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:pdex_PDB_model_num>1
</PDBx:pdex_PDB_model_num>
<PDBx:type_symbol>N</PDBx:type_symbol>
</PDBx:atom_site>
```

```
<PDBx:atom_site id="4383">
```

```
...
<PDBx:auth_asym_id>A</PDBx:auth_asym_id>
<PDBx:auth_atom_id>FE</PDBx:auth_atom_id>
<PDBx:auth_comp_id>HEM</PDBx:auth_comp_id>
<PDBx:auth_seq_id>142</PDBx:auth_seq_id>
<PDBx:group_PDB>HETATM</PDBx:group_PDB>
<PDBx:label_asym_id>E</PDBx:label_asym_id>
<PDBx:label_atom_id>FE</PDBx:label_atom_id>
<PDBx:label_comp_id>HEM</PDBx:label_comp_id>
<PDBx:label_entity_id>3</PDBx:label_entity_id>
<PDBx:label_seq_id></PDBx:label_seq_id>
<PDBx:occupancy>1.00</PDBx:occupancy>
<PDBx:pdex_PDB_model_num>1
</PDBx:pdex_PDB_model_num>
<PDBx:type_symbol>FE</PDBx:type_symbol>
</PDBx:atom_site>
```

30

chainと分子の対応

struct_asymCategoryと entityCategoryの対応

```

<PDBx:entityCategory>
  <PDBx:entity_id="1">
    <PDBx:formula_weight>15150.497</PDBx:formula_weight>
    <PDBx:pdxb_description>HEMOGLOBIN (ALPHA CHAIN)</PDBx:pdxb_description>
    <PDBx:pdxb_number_of_molecules>2.</PDBx:pdxb_number_of_molecules>
    <PDBx:src_method>man</PDBx:src_method>
    <PDBx:type>polymer</PDBx:type>
  </PDBx:entity>
  <PDBx:entity_id="2">
    <PDBx:formula_weight>15899.374</PDBx:formula_weight>
    <PDBx:pdxb_description>HEMOGLOBIN (BETA CHAIN)</PDBx:pdxb_description>
    <PDBx:pdxb_mutation>V1M, W37Y</PDBx:pdxb_mutation>
    <PDBx:pdxb_number_of_molecules>2.</PDBx:pdxb_number_of_molecules>
    <PDBx:type>polymer</PDBx:type>
  </PDBx:entity>
</PDBx:entityCategory>
...
<PDBx:struct_asymCategory>
  <PDBx:struct_asym id="A">
    <PDBx:entity_id>1</PDBx:entity_id>
  </PDBx:struct_asym>
  <PDBx:struct_asym id="B">
    <PDBx:entity_id>2</PDBx:entity_id>
  </PDBx:struct_asym>
  <PDBx:struct_asym id="C">
    <PDBx:entity_id>1</PDBx:entity_id>
  </PDBx:struct_asym>
  <PDBx:struct_asym id="D">
    <PDBx:entity_id>2</PDBx:entity_id>
  </PDBx:struct_asym>
</PDBx:struct_asymCategory>

```

31

アンケートから(2)

- 「“MSE”といったアミノ酸表記を“MET”に直した場合のXMLファイルや、少なくとも対応付けリストがあると助かります。」

→ mmCIF/PDBMLのpdxb_struct_mod_residueカテゴリに対応付けリストがある！

```

<PDBx:pdxb_struct_mod_residueCategory>
  <PDBx:pdxb_struct_mod_residue id="1">
    <PDBx:PDB_ins_code xsi:nil="true" />
    <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
    <PDBx:auth_comp_id>MSE</PDBx:auth_comp_id>
    <PDBx:auth_seq_id>45</PDBx:auth_seq_id>
    <PDBx:details>SELENOMETHIONINE</PDBx:details>
    <PDBx:label_asym_id>A</PDBx:label_asym_id>
    <PDBx:label_comp_id>MSE</PDBx:label_comp_id>
    <PDBx:label_seq_id>45</PDBx:label_seq_id>
    <PDBx:parent_comp_id>MET</PDBx:parent_comp_id>
  </PDBx:pdxb_struct_mod_residue>

```

32

演習課題1

- エントリー「101d」を検索して、以下のことを調べる。
 - 発表された論文は？
 - 実験手法は？
 - 含まれる分子種は？
 - 解像度(もしあれば)は？

ヒント: citation, exptl, entity, refine

33

演習課題2

- キーワード検索で「aldehyde dehydrogenase」を入力にして、以下のことを調べる。
 - 何件のエントリーがヒットするか？
 - Release date が最も古いエントリーはどれか？
 - その論文は何年に発表されたか？
 - そのエントリーは何年にPDBからリリースされたか？
 - その由来する生物種は何か？
 - その活性残基はどれか？

ヒント: citation, database_PDB_rev, entity_src_gen, struct_site

34

演習課題3

- PDBフォーマットとmmCIFを比べてみる。エントリー「1gof」を例にとる。
 - PDB, mmCIFで、release dateはそれぞれどこに書かれているか？
 - PDB, mmCIFで、登録者名はそれぞれどこに書かれているか？
 - PDB, mmCIFで、実験手法はそれぞれどこに書かれているか？
 - PDB, mmCIFで、由来する生物種(biological species)はそれぞれどこに書かれているか？
 - PDBの”ATOM”行はmmCIFのどこにどう対応するか？

ヒント: database_PDB_rev, audit_author, exptl, entity_src_gen, atom_site

35

演習課題4

- mmCIFとPDBMLを比較してみる。
 - 演習課題3と同じことをmmCIFとPDBMLで行う。
 - http://service.pdbj.org/mine/pdbml_image.html も参照。
 - PDBMLには原子座標の部分だけを簡略化した「extatom」と呼ばれる形式もある。PDBjのサイトで、これとmmCIFと対応する部分を比較せよ。

ヒント: database_PDB_rev, audit_author, exptl, entity_src_gen, atom_site

36

演習課題5

- PDBjのトップページから、今週の更新情報を調べる。
 - 今週新規に追加されたエントリーは何件あるか？
 - いくつか例を眺めてみる。
 - 今週更新されたエントリーは何件あるか？
 - いくつか例を眺めてみて、何度更新されたかを調べる。
 - “database_PDB_rev” カテゴリをみるとわかる。
 - どこが更新されたかを調べる。
 - “database_PDB_rev_record”カテゴリをみると(だいたい)わかる。

37

演習課題6

- Advanced Searchを用いて以下のことを調べる。
 - Release dateが2000年1月1日以降のエントリーは何件あるか？
 - Release dateが1999年12月31日以前のエントリーは何件あるか？

38

演習課題7

- Advanced Searchを使って、引用文献に関して調べてみる。
 - “Walker, J.E.” を著者に含むエントリーは何件あるか？
 - ヒットしたものの中には、一見して “Walker, J.E.” を含まないように見えるものもある。なぜか？
 - “Primary citation only!” ボタンをチェックして、同じ検索を試してみる。結果はどう変わるか？
 - 雑誌 “Nature” に掲載されたエントリーは何件あるか？
 - 雑誌 “Science” に2005年に掲載されたエントリーはいくつあるか？

ヒント: audit_author, citation, citation_author

39

演習課題8

- Advanced Searchをつかう。
 - “polypeptide(L)” (要するに普通の蛋白質) を含むエントリーは何件あるか？
 - “polypeptide(D)” のみを含むエントリーを検索するにはどうすればよいか？
 - 蛋白質とDNAの複合体を検索するにはどうすればよいか？
 - 蛋白質、DNA、RNAを含むエントリーは何件あるか？
 - RNAと “polysaccharide(D)” の両方を含むエントリーはあるか？
 - その他色々な組み合わせを試してみる。

ヒント: entity_poly, entity

40

演習課題9

- Advanced Searchを使って蛋白質の名前で検索する。
 - “Compound information” に “prion” を入れて検索する。
 - ヒットしたエントリーのサマリーページをいくつか眺めて、“prion” という単語がどこに現れているかを観察する。
 - mmCIFまたはPDBML-noatomを読んで、prionという単語がどこに含まれているかを調べる。(どのタグか?)
 - “prion”の代わりに “oxidoreductase”として同じことを試してみる。

ヒント: entity, struct, pdbx_database_related, citation, entity_name_com, struct_keywords

演習課題10

- Advanced Search
 - “Title” に “crystal” と入力して検索。
 - 前問と同じように、サマリーページと元ファイルを確認する。
 - 他の単語でも試してみる。
 - たとえば “DNA complex”

演習課題11

- PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>) にアクセスして、“Borhani DW [au] truncated human apolipoprotein 1997” を検索する。
 - PubMed ID (PMID) をコピーする。
- Advanced Searchで “Other DB”から PubMedを選び、“ID:” の箱にPMIDをペーストして、検索する。
 - PDB IDは何か？
 - Primary citation とPubMedでの検索結果を比較する。
 - PMIDはmmCIF/PDBMLのどこに記載されているか？

ヒント: citation

43

演習課題12

- Advanced Search の “Other DB”から “EC Number” (酵素番号)を選択する。
- “ID:”の箱に “1.1.1.1” (alcohol dehydrogenase)を入力する。
 - ヒットするエントリーは何件か？
 - EC番号は mmCIF/PDBMLのどこに書かれているか？
 - ついでにいくつかのエントリーで活性残基も調べてみる。

ヒント: entity, struct_site

44

演習課題13

- Advanced Search の “Other DB” から “GO” (gene ontology) を選択して、“ID:” には何も入れずに検索してみる。
 - 何件ヒットするか？
 - いくつかのエントリーで “Functional Details” ページを眺めてみる。
 - GO の ID は mmCIF/PDBML のどこに書かれているか？

ヒント: gene_ontology (PDBMLplusのみ)

45

演習課題14

- Advanced Search の “Other DB” から “UniProt” を選択して、“ID:” には何も入れずに検索する。
 - これでヒットするエントリーの例えば 101m のサマリーページを見ると、ページ下の方に UniProt へのリンクがある。これをたどって、この蛋白質のアノテーションを調べてみる。
 - “Functional Details” ページ (PDBj) に gene ontology のアノテーションはあるか？
 - 対応する UniProt エントリー (リンク先) には gene ontology のアノテーションがあるか？
 - UniProt の ID (P02185 など) は mmCIF/PDBML のどこに書かれているか？

ヒント: gene_ontology(PDBMLplus only), struct_ref

46

演習課題15

- 前問と同様のことを “GB” (GenBank) , “EMBL” および “PIR”でもやってみる。
 - それぞれで、最も最近リリースされたエントリーは何か？
 - (PIRは2002年にUniProtに併合された)

ヒント: gene_ontology(PDBMLplus only), struct_ref

47

演習課題16

- Advanced Search の “Ligands and Prosthetic groups” に “ATP” と入力して検索する。
 - ヒットしたエントリーのいくつかをサマリーページで確認し、jV3でグラフィクス表示してどこに ATP が結合しているかを確認する。
 - mmCIF/PDBMLではどこに “ATP”に関する記述があるか？
 - PDBフォーマットファイルではどうか？
 - ATPの原子座標もみってみる。蛋白質の原子座標の表記とどのように異なるか？

ヒント: chem_comp, entity, pdbx_entity_nonpoly

48

演習課題17

- 前問と同様のことを色々なリガンド名で試してみる: FAD, NAD, HEM, PO4 など。
- PDB に含まれるリガンド化合物は chemical component dictionary で定義されている。詳しくは、
 - <http://www.wwpdb.org/ccd.html>

49

演習課題18

- Advanced Searchの “Numbers of chains and Chain length” でポリマー分子の数(上限、下限)とポリマー分子の長さ(上限、下限)をいろいろ指定して検索してみる。
 - ポリマーは蛋白質、核酸などを含む。
 - 例えば、最低(min)10分子のポリマーを含み、長さが最大20ユニット(残基、塩基)のポリマーを含むエントリーは？
 - ポリマーとしては蛋白質のみを含み、かつ二量体として解かれた構造はどうやって検索するか？

ヒント: entity, entity_poly, entity_poly_seq, pdbx_poly_seq_scheme

50

演習課題19

- Advanced Searchの“Experimental Technique”にはどのような選択肢があるか？
 - それぞれを選んで検索し、何件ヒットするか確認する。
 - この選択肢以外に登録されているデータは(今のところ)ない。
 - 検索結果の絵をぱっと見て、実験手法ごとにどのような傾向があるか考えてみる。
 - 実験手法の情報はmmCIF/PDBMLのどこに記載されているか？
 - 次に resolution (解像度)も指定してみる。
 - 実験手法によっては解像度がないものもあることに注意。
 - ところが、“SOLUTION NMR”で解像度 2.5Å以下と指定するとヒットするものがある。なぜか？

ヒント: exptl, refine, reflns

51

演習課題20

- Advanced Searchの“Source”に生物種名を入れて検索する。(その蛋白質などが由来する生物種)
 - 例: human, bos taurus, pyrococcusなど
- “Host species”に生物種名を入れて検索する。(その蛋白質などを精製するのに用いた生物種)
- これらの情報はmmCIF/PDBMLのどこに記載されているか？

ヒント: entity_src_gen, entity_src_nat, entity

52