Bioinformatics III

Analysis and prediction of 3D macromolecule structures

Lecture 3 – structural file formats (CIF)

Saulius Gražulis
2021 m.

CIF an mmCIF formats

- ASCII (CIF 2: UTF-8) encoded files
- Free format syntax
- Data identified by keywords
- Uses relational data model
- Each *data value* is associated with a *data name,* forming a *data item*
- Meanings of data names are specified in CIF dictionaries

http://www.iucr.org/iucr-top/cif/standard/cifstd1.html
http://www.iucr.org/iucr-top/cif/spec/version1.1/cifsyntax.html

Example of a CIF file

```
data_1KNV
#
_entry.id   1KNV
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    1.044
...
_cell.entry_id           1KNV
_cell.length_a           121.230
_cell.length_b           122.280
_cell.length_c           56.870
_cell.angle_alpha        90.00
_cell.angle_beta         90.00
_cell.angle_gamma        90.00
...
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
...
ATOM   1    N  N   . ASN A 1 4   ? 3.407  40.303 50.109  1.00 66.19 ? ? ? ? ? 4    ASN A N   1
ATOM   2    C  CA  .
ASN A 1 4   ?  4.752
              40.029 49.523  1.00 67.25 ? ? ? ? ? 4    ASN A CA  1
```

Languages (Computer Sci.)

In CS, a *language* L is a pair (A, W $\subset$ A*), where:

A is a *finite* alphabet (i.e. a *finite* set of symbols),

A* is the (infinite) set of all possible *strings* over A

W is a *subset* of A*.

Grammars

Example of a **grammar**:

$$R \to S \mid S + R \mid S - R$$
$$S \to D \mid D \times S \mid D / S$$
$$D \to V \mid (R)$$
$$V \to a \mid b \mid c$$

Correct (derivable) sentence:

$$(a+b)/(a-b\times b/c)$$

Incorrect (non-derivable) sentence:

$$((a)(+-b)/(()a-b\,bbcdef\times b/c)$$

Backus–Naur Form (BNF)

```
<expression> ::= <product>
               | <product> + <expression>
               | <product> - <expression>

<product>  ::= <term>
               | <term> * <product>
               | <term> / <product>

<term> ::= <identifier> | ( <expression> )

<identifier> ::= a | b | c
```

CIF and STAR syntax

CIF grammar in Backus-Naur form:

```
...
<data_block>   ::= <data_heading> <data>+ { <wspace>+ | <EOF> }
<data_heading> ::= <DATA_> <non_blank_char>+
<data>         ::= { <wspace>+ <data_name> <wspace>* <blank>
                     <data_value_1> }
                 | { <wspace>+ <data_name> <wspace>* <terminate>
                     <data_value_2> }
                 | <data_loop>

<data_loop>    ::= <wspace>+ <LOOP_> <data_loop_field> <data_loop_values>

<data_loop_field>   ::= { <wspace>+ <data_name> }+
<data_name>         ::= '_' <non_blank_char>+
<data_loop_values>  ::= { { <wspace>* <blank> <data_value_1> }
                        | { <wspace>* <terminate> <data_value_2> } }+
...
```

CIF 1: http://ww1.iucr.org/iucr-top/cif/spec/version1.1/cifsyntax.html#gram
CIF 2: Bernstein 2016 https://doi.org/10.1107/s160057671502187l
CIF 2 grammar: https://journals.iucr.org/j/issues/2016/01/00/aj5269/aj5269sup1.txt
GitHub: https://github.com/COMCIFS

CIF syntax features

```
# Comments start with a “hash” (#) symbol and extend to the end of the line
# CIF 1: only ASCII symbols are allowed; CIF 2: uses UTF-8 encoding

data_DataName
_tag1 value   # values without spaces or quotes can be specified as they are
_tag2 1.23(3) # Numbers carry optional precision (standard uncertainty, su)
_tag3 'values with spaces must be in single ...'
_tag4 "... or double quotes -- this is a quoted string (q.s.)"
_tag5 'a word like d'Alamber with a quote may be in the middle of a q.s.(!)'
_tag5a
# Comments may be inserted where the white space is allowed
'a value may be anywhere in the file, also on another line'

loop_
_tag6 # Data tables (loops) MAY be arbitrarily split into lines.
_tag7 _tag8
123 456 789
111 222
333

DaTa_NextDataName # CIF keywords are case insensitive, but values are
_tag1 123 # Data names MUST be unique within the data block
          # but may be repeated in subsequent blocks

# No special mark at the end of the file or data stream
```

CIF semantics; CIF dictionaries

Q: what does '_atom_site_label' mean?
What data names are used for coordinates?

```
data_atom_site_fract_
    loop_ _name                  '_atom_site_fract_x'
                                 '_atom_site_fract_y'
                                 '_atom_site_fract_z'
    _category                     atom_site
    _type                         numb
...
    _list_reference              '_atom_site_label'
    _definition
;               Atom-site coordinates as fractions of the _cell_length_ values.
;

data_atom_site_label
    _name                        '_atom_site_label'
    _category                     atom_site
    _type                         char
...
_definition
;               The _atom_site_label is a unique identifier for a particular site
                in the crystal.
...
```

CIF 1: https://github.com/COMCIFS/DDL1-legacy-dictionaries
CIF 2: https://github.com/COMCIFS
CIF Core: https://github.com/COMCIFS/cif_core/blob/master/cif_core.dic

CIF “dictionaries of dictionaries”, DDL

What does '_name' mean?
Which data name is used to specify value type?

```
data_name
    _definition
;               The data name(s) of the defined item(s). If data items are
                closely related or represent an irreducible set, their names
                may be declared as a looped sequence in the same definition.
;
    _name                       '_name'
    _category                     name
    _type                         char
    _list                         both
    loop_ _example              '_atom_site_label'
                                '_atom_attach_all   _atom_attach_ring'
```

DDL1: https://www.iucr.org/resources/cif/ddl/ddl1
DDL2: https://www.iucr.org/resources/cif/ddl/ddl2
DDLm: https://www.iucr.org/resources/cif/ddl/ddlm

mmCIF, DDL2 dictionary

The original CIF DDL1 dictionary did **not** have means to describe macromolecules

PDB created mmCIF (macromolecular CIF) format (syntactically compatible with CIF) and the DDL2 dictionary

Advantages of the (mm)CIF format

- Plain text (ASCII or UTF-8), human and machine readable
- Formally defined syntax
- Machine readable semantics (dictionaries)
- Applicable to all kinds of information
- Easy means to add new data items

Drawbacks of the (mm)CIF format

- Complex grammar, needs non-trivial parsers
- Many common errors are difficult to localise or even to detect (e.g. missing 'loop_' elements)
- Some aspects of semantics are only interpretable by humans
- Multiple slightly different dialects

PDB XML schema

mmCIF dictionaries can be converted into an XML schema

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock  datablockName="1KNV"
        xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx.xsd"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://deposit.pdb.org/pdbML/pdbx.xsd pdbx.xsd">
...
  <PDBx:atom_siteCategory>
    <PDBx:atom_site id="1">
      <PDBx:group_PDB>ATOM</PDBx:group_PDB>
      <PDBx:type_symbol>N</PDBx:type_symbol>
      <PDBx:label_atom_id>N</PDBx:label_atom_id>
      <PDBx:label_alt_id xsi:nil="true" />
      <PDBx:label_comp_id>ASN</PDBx:label_comp_id>
      <PDBx:label_asym_id>A</PDBx:label_asym_id>
      <PDBx:label_entity_id>1</PDBx:label_entity_id>
      <PDBx:label_seq_id>4</PDBx:label_seq_id>
      <PDBx:Cartn_x>3.407</PDBx:Cartn_x>
      <PDBx:Cartn_y>40.303</PDBx:Cartn_y>
      <PDBx:Cartn_z>50.109</PDBx:Cartn_z>
...
```

Ontologies and semantic networks

Ontology (Greek ων „exist“, λόγος „word“, „notion“) — a branch of philosophy that examines the question “what exists?”.

Ontologies (pl.) — in Computer Science, a formal description of terms and their logical relations in some application domain.

https://en.wikipedia.org/wiki/Ontology
https://en.wikipedia.org/wiki/Ontology_(information_science)
http://lt.wikipedia.org/wiki/Ontologija
http://lt.wikipedia.org/wiki/Ontologija_(informatika)

“Ideal” format?

Text, stadard encoding ASCII -> UTF8

Record <=> line

Space separated fields

Keywords indicate record types

Fixed record fields and types?

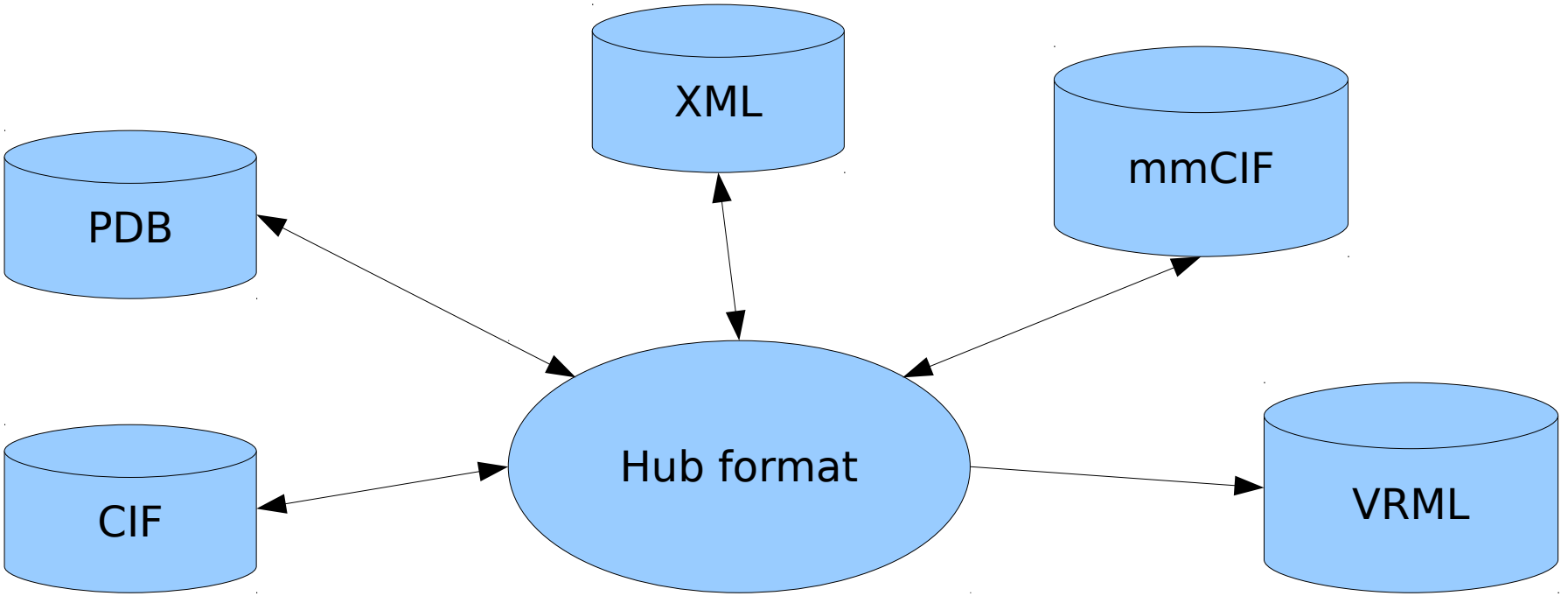
No size restrictions!

Example ...

```
FORMAT My ideal macromolecular data format ver. 0.0
#
# Komentarai gali būti skirti žmogui
#
TITLE   Restrikcijos endonukleazės struktūra
AUTHORS Saulius Gražulis; Elena Manakova (Манакова, Елена)
CELL    100.0 100.0 100.0 100.0 90 90 90
SPACEGROUP P212121
#
ATOM N  ASN A 4  3.407(1)  40.303(2)  50.109(11) 1.00 66.19  N
ATOM CA ASN A 4  4.752     40.029     49.523     1.00 67.25  C
...
```

Possible uses of the format

Hub formats



Example of a successful Hub Format: netpbm
http://netpbm.sourceforge.net/