Data File Formats

Introduction

- There are dozens of file formats for chemical data.
  - We will do an overview of a few that are often used in structural bioinformatics.

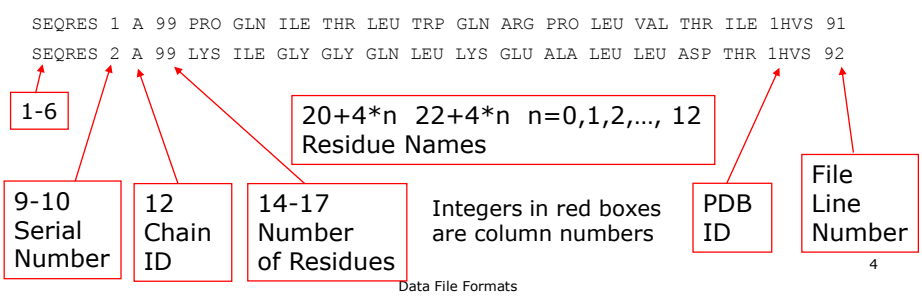
PDB File Format (1)

- The PDB file format specification is rather extensive.
  - A tutorial on the topic can be found at: http://www.wwpdb.org/documentation/format23/v2.3.html
- A PDB file includes the following information:
  - HEADER
    - First line of the file, contains PDB ID code, classification, and date of deposition.
  - COMPND
    - A description of the macromolecular contents.
  - SOURCE
    - Biological source of the macromolecule.
  - REVDAT
    - Revision dates.
  - JRNL
    - Reference to journal article dealing with this structure.

3

PDB File Format (2)

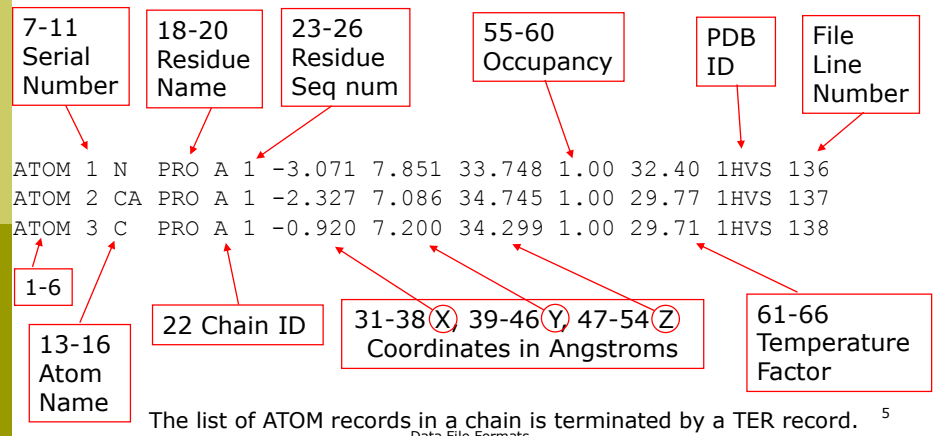
- A PDB file information continued:
  - AUTHOR
    - List of contributors.
  - REMARK
    - General remarks, some structured, some free form.
  - SEQRES
    - Primary sequence of the residues on a protein backbone.



4

PDB File Format (3)

- For your assignment the most important type of record is that ATOM record:
  - Contains the coordinates of a particular atom.



The list of ATOM records in a chain is terminated by a TER record.

PDB File Format (4)

- Other record types:
  - HET
    - Names of hetero-atoms
      - Usually designates atoms and molecules that are separate from the protein.
  - FORMUL
    - Chemical formula for the hetero-atoms.
  - HELIX
    - Location of helices (designated by depositor).
  - SHEET
    - Location of beta sheets (designated by depositor).
  - SSBOND
    - Location of disulfide bonds (if they exist).
  - ORIGn
    - Scaling factors to transform from orthogonal coordinates.
  - SCALEn
    - Scaling factors transforming to fractional crystal coordinates.

A New File Format: mmCIF

- Macromolecular Crystallographic Information File
  - Derived from CIF
    - S.R. Hall, F.H. Allen, and I.D. Brown. 1991. The Crystallographic Information File (CIF): a new standard archive file for crystallography. Acta Cryst. (1991). A47, 655-685.
  - Dictionary based
    - The macromolecular CIF dictionary is a set of data names designed to describe the macromolecular crystallographic experiment and its results.
    - Each data item in a mmCIF matches an entry in the dictionary.
    - This enables the validation of data and gives a more consistent representation of structures.
  - Easier to parse with less ambiguity.

Comparing PDB Files and mmCIF (1)

- PDB Header:

HEADER HYDROLASE(ACID PROTEASE) 26-JAN-94 1HVI

becomes:

```
_struct.entry_id 1HVI
_struct.title
;INFLUENCE OF STEREOCHEMISTRY ON ACTIVITY AND BINDING MODES FOR C2
   SYMMETRY-BASED DIOL INHIBITORS OF HIV-1 PROTEASE ;
_struct_keywords.entry_id 1HVI
_struct_keywords.pdbx_keywords 'HYDROLASE(ACID PROTEASE)'
_struct_keywords.text 'HYDROLASE(ACID PROTEASE)'
```

Comparing PDB and mmCIF (2)

- The equivalent of ATOM records:

```
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.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
```


Comparing PDB and mmCIF (3)

- The equivalent of ATOM records (cont.):

```
ATOM 1 N N . PRO A 1 1 ? -3.609 7.549 33.926 1.00 27.71 ? ? ? ? ?
   1 PRO A N 1
ATOM 2 C CA . PRO A 1 1 ? -2.655 6.720 34.710 1.00 27.34 ? ? ? ? ?
   1 PRO A CA 1
ATOM 4 O O . PRO A 1 1 ? -1.000 7.590 33.270 1.00 29.50 ? ? ? ? ?
   1 PRO A O 1
```

...

PDB records for these first three atoms:

```
ATOM      1  N   PRO A   1      -3.609   7.549  33.926  1.00 27.71
ATOM      2  CA  PRO A   1      -2.655   6.720  34.710  1.00 27.34
ATOM      3  C   PRO A   1      -1.219   6.969  34.289  1.00 27.56
```

Comparing PDB and mmCIF (4)

- To see a good tutorial on mmCIF you can visit the SDSC (San Diego Supercomputer Center):
  - http://www.sdsc.edu/pb/cif/tutorial_mm.html
  - This includes many examples of how dictionary entries and their associated values appear in a typical mmCIF file.
    - There are also several interesting links from their homepage.

PubChem File Formats: ASN.1 (1)

- ASN.1 (Abstract Syntax Notation 1):
  - A standardized data specification language used to represent complex data types.
    - Data structure is hierarchical
    - Designed by Xerox
  - A formal notation used for describing data transmitted by telecommunications protocols and physical representation of such data.
  - Also used in telephone systems, air traffic, building and machine control, toll highways, smart cards, security and much more.
  - Used by NCBI to store GenBank, PubMed, MMDB.

National Center for Biotechnology Information

PubChem File Formats: ASN.1 (2)

- Abstract Syntax Notation 1 for water:

```
PC-Compound ::= {
  id {
    id cid 962
  },
  atoms {
    aid {
      1,
      2,
      3
    },
    element {
      o,
      h,
      h
    }
  },
  bonds {
    aid1 {
      1,
      1
    },
    aid2 {
      2,
      3
    },
    order {
      single,
      single
    }
  },
```

```
  coords {
    {
      type {
        twod,
        computed,
        units-unknown
      },
      aid {
        1,
        2,
        3
      },
      conformers {
        {
          x {
            { 25369358062744, 10, -13 },
            { 30738716125488, 10, -13 },
            { 2, 10, 0 }
          },
          y {
            { -155000001192093, 10, -15 },
            { 155000001192093, 10, -15 },
            { 155000001192093, 10, -15 }
          }
        }
      }
    }
  },
  charge 0,
```

```
  props {
    {
      urn {
        label "Count",
        name "Hydrogen Bond Acceptor",
        datatype uint,
        implementation "E_NHACCEPTORS",
        version "3.328",
        software "Cactvs",
        source "xemistry.com",
        release "2006.10.23"
      },
      value ival 1
    },
    {
      urn {
        label "Count",
        name "Hydrogen Bond Donor",
        datatype uint,
        implementation "E_NHDONORS",
        version "3.328",
        software "Cactvs",
        source "xemistry.com",
        release "2006.10.23"
      },
      value ival 1
    },
```

ETC. …

PubChem File Formats: XML (1)

- Extensible Markup Language
  - Is a general-purpose markup language that supports a wide variety of applications.
  - Just like HTML it is also built around an hierarchical layout that uses markup tags.
  - It is quite versatile.
  - Almost human readable.
    - However, files tend to be rather long.
    - The PubChem entry for water has an XML file that is 406 lines long!
      - The first few lines of this file appear next:

PubChem File Formats: XML (2)

- Water in XML:

```
<?xml version="1.0" ?>
- <PC-Compound xmlns="http://www.ncbi.nlm.nih.gov"
    xmlns:xs="http://www.w3.org/2001/XMLSchema-instance"
    xs:schemaLocation="http://www.ncbi.nlm.nih.gov
    ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.xsd">
- <PC-Compound_id>
- <PC-CompoundType>
- <PC-CompoundType_id>
  <PC-CompoundType_id_cid>962</PC-CompoundType_id_cid>
    </PC-CompoundType_id>
    </PC-CompoundType>
    </PC-Compound_id>
- <PC-Compound_atoms>
- <PC-Atoms>
- <PC-Atoms_aid>
  <PC-Atoms_aid_E>1</PC-Atoms_aid_E>
  <PC-Atoms_aid_E>2</PC-Atoms_aid_E>
  <PC-Atoms_aid_E>3</PC-Atoms_aid_E>
    </PC-Atoms_aid>
- <PC-Atoms_element>
  <PC-Element value="o">8</PC-Element>
  <PC-Element value="h">1</PC-Element>
  <PC-Element value="h">1</PC-Element>
    </PC-Atoms_element>
    </PC-Atoms>
    </PC-Compound_atoms>
```


PubChem File Formats: XML (3)

- Water in XML (cont.):

```
- <PC-Compound_bonds>
- <PC-Bonds>
- <PC-Bonds_aid1>
  <PC-Bonds_aid1_E>1</PC-Bonds_aid1_E>
  <PC-Bonds_aid1_E>1</PC-Bonds_aid1_E>
    </PC-Bonds_aid1>
- <PC-Bonds_aid2>
  <PC-Bonds_aid2_E>2</PC-Bonds_aid2_E>
  <PC-Bonds_aid2_E>3</PC-Bonds_aid2_E>
    </PC-Bonds_aid2>
- <PC-Bonds_order>
  <PC-BondType value="single">1</PC-BondType>
  <PC-BondType value="single">1</PC-BondType>
    </PC-Bonds_order>
    </PC-Bonds>
    </PC-Compound_bonds>
- <PC-Compound_coords>
- <PC-Coordinates>
- <PC-Coordinates_type>
  <PC-CoordinateType value="twod">1</PC-CoordinateType>
  <PC-CoordinateType value="computed">5</PC-CoordinateType>
  <PC-CoordinateType value="units-unknown">255</PC-CoordinateType>
    </PC-Coordinates_type>
```


357 lines to go…

PubChem File Formats: SDF (1)

- SDF is a chemical file format
  - File extensions: .mol .sd .sdf
  - A **MDL Molfile** is a file format designed by MDL, for storing data about the atoms, bonds, connectivity and atom coordinates of a molecule.
  - The molfile consists of header information, the Connection Table (CT) containing atom information, bond connections and types, followed by records for more complex information.

The format is owned by MDL (Molecular Design Limited).

PubChem File Formats: SDF (2)

- SDF for glycine (CID 750):

```
750
  -OEChem-01310721062D

 10  9  0     0  0  0  0  0  0999 V2000
    2.5369    0.7500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.4030   -0.7500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350    0.2500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    4.2690    0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.4030    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.6675    1.2249    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    3.8705    1.2249    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    5.6719    0.5600    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350   -0.3700    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.0000    0.4400    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  5  1  0  0  0  0
  1 10  1  0  0  0  0
  2  5  2  0  0  0  0
  3  4  1  0  0  0  0
  3  8  1  0  0  0  0
  3  9  1  0  0  0  0
  4  5  1  0  0  0  0
  4  6  1  0  0  0  0
  4  7  1  0  0  0  0
M  END
```

PubChem File Formats: SDF (3)

- SDF for glycine (CID 750)
  - Other useful information:

```
<PUBCHEM_CACTVS_HBOND_ACCEPTOR>
3
> <PUBCHEM_CACTVS_HBOND_DONOR>
2
> <PUBCHEM_CACTVS_ROTATABLE_BOND>
1
> <PUBCHEM_IUPAC_NAME>
2-aminoacetic acid
> <PUBCHEM_NIST_INCHI>
InChI=1/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)/f/h4H
> <PUBCHEM_CACTVS_XLOGP>
-3.4
> <PUBCHEM_OPENEYE_MF>
C2H5NO2
> <PUBCHEM_OPENEYE_MW>
75.0666
> <PUBCHEM_OPENEYE_CAN_SMILES> C(C(=O)O)N
```