Bioinformatics III

Analysis and prediction of 3D macromolecule structures

Lecture 2 – structural file formats (PDB)

Saulius Gražulis
2021 m.

“To me, you understand something only if you can program it. (You, not someone else!)”

Gregory Chaitin, “Meta Math! -- The Quest for Omega Ω” //Vintage Books, A Division of Random House, Inc., New York, First edition (2006), chapter “Preface”, page xiii.

Structural data

- Atomic coordinates
- Temperature (B) factors, occupancies
- Crystallographic information (unit cell parameters, symmetry information)
- Data and structure quality information
- Additional (meta)data
  - macromolecule sequence
  - secondary structure assignment
  - biochemical and biological data ...

PDB file format

- ASCII encoded text files
- Fixed column format
- One record – one line of the file
- Each record starts with a keyword

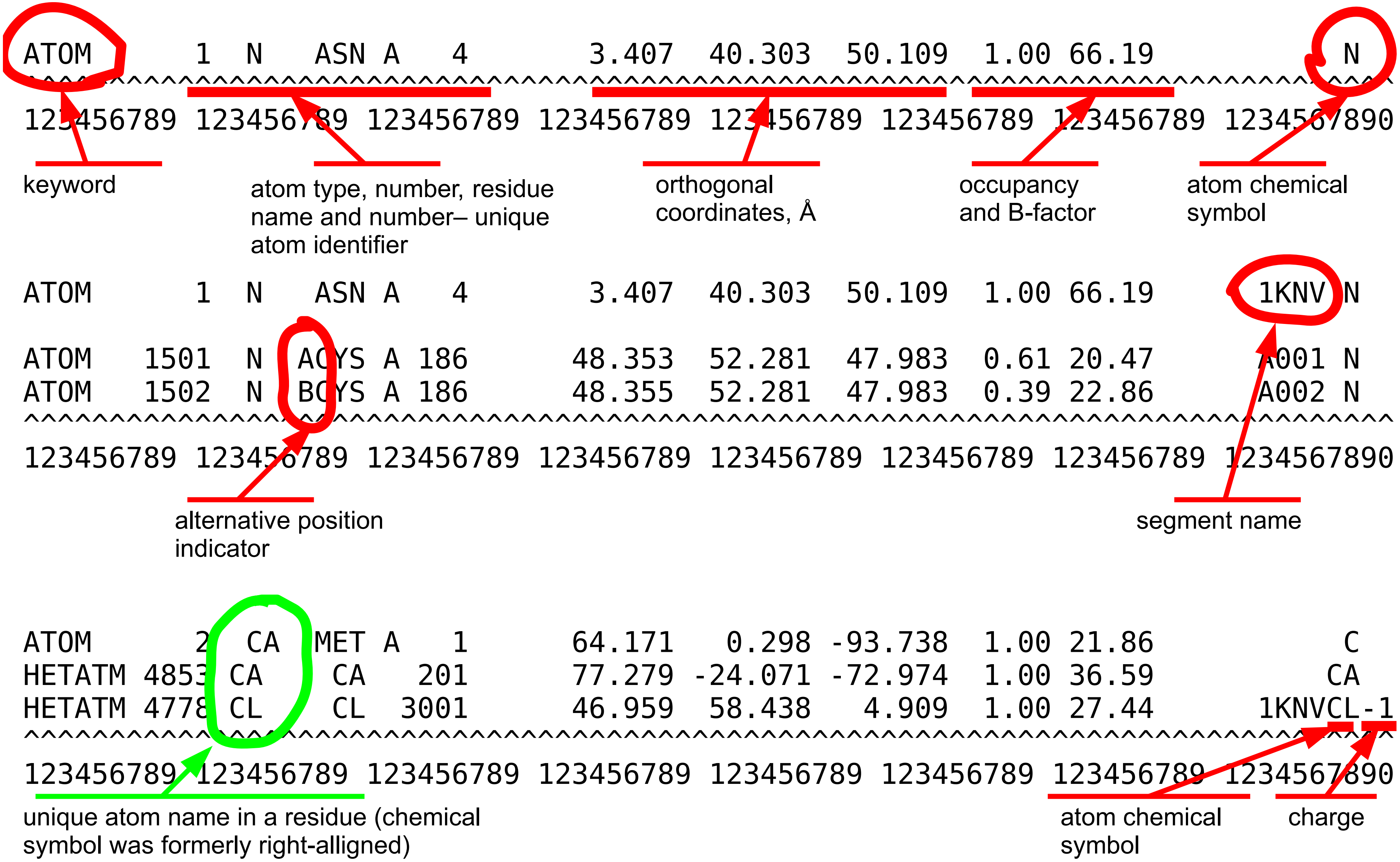
http://www.wwpdb.org/docs.html
http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html
ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_A4.pdf

Example of a PDB file

```
HEADER    HYDROLASE                               15-SEP-05   2C1L
TITLE     STRUCTURE OF THE BFII RESTRICTION ENDONUCLEASE
...
JRNL        AUTH   S.GRAZULIS,E.MANAKOVA,M.ROESSLE,M.BOCHTLER,
JRNL        AUTH 2 G.TAMULAITIENE,R.HUBER,V.SIKSNYS
...
REMARK   2 RESOLUTION. 1.90 ANGSTROMS.
...
CRYST1  138.925  138.925   94.135  90.00  90.00  90.00 I 4          16
SCALE1      0.007198  0.000000  0.000000        0.00000
SCALE2      0.000000  0.007198  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010623        0.00000
ATOM      1  N  AMET A   1      40.881   1.095  49.888  0.33 24.33           N
ATOM      2  N  BMET A   1      40.265   1.169  49.581  0.33 24.33           N
...
END
```

PDB format ATOM records

From the original PDB 1KNV and 2EZV entries:



```
ATOM      1  N   ASN A   4       3.407  40.303  50.109  1.00 66.19           N
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
123456789 123456789 123456789 123456789 123456789 123456789 123456789 1234567890
```

keyword

atom type, number, residue name and number– unique atom identifier

orthogonal coordinates, Å

occupancy and B-factor

atom chemical symbol

```
ATOM      1  N   ASN A   4       3.407  40.303  50.109  1.00 66.19      1KNV N

ATOM   1501  N  ACYS A 186      48.353  52.281  47.983  0.61 20.47      A001 N
ATOM   1502  N  BCYS A 186      48.355  52.281  47.983  0.39 22.86      A002 N
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
123456789 123456789 123456789 123456789 123456789 123456789 123456789 1234567890
```

alternative position indicator

segment name

```
ATOM      2  CA  MET A   1      64.171   0.298 -93.738  1.00 21.86           C
HETATM 4853 CA    CA   201      77.279 -24.071 -72.974  1.00 36.59          CA
HETATM 4778 CL    CL  3001      46.959  58.438   4.909  1.00 27.44      1KNVCL-1
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
123456789 123456789 123456789 123456789 123456789 123456789 123456789 1234567890
```

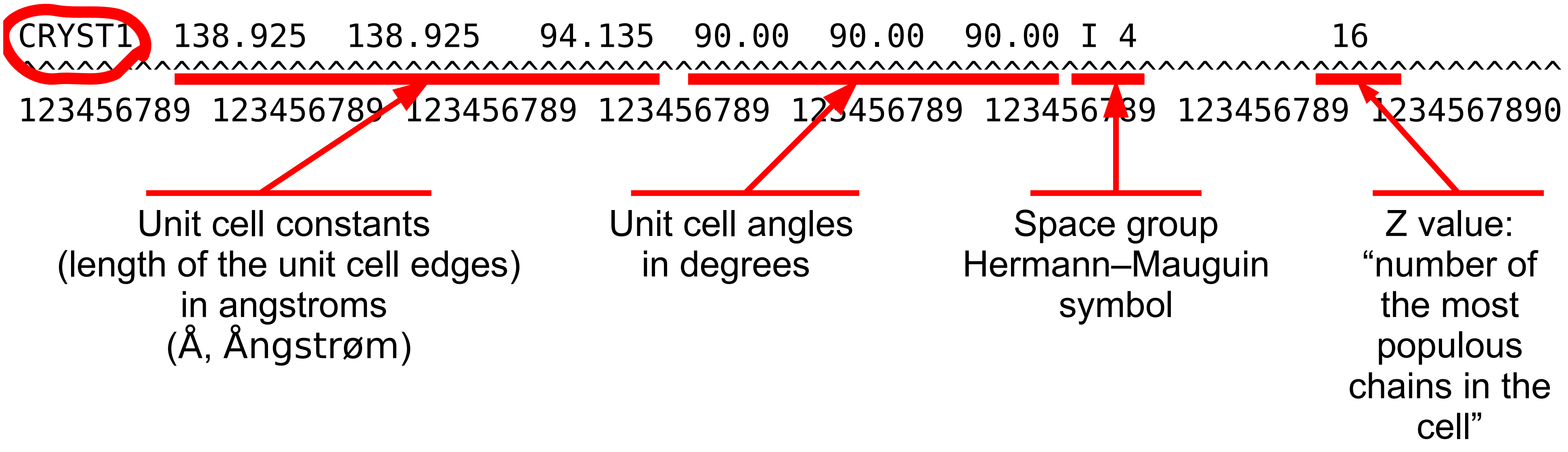
unique atom name in a residue (chemical symbol was formerly right-alligned)

atom chemical symbol

charge

Crystallographic information in a PDB file – CRYST1 record

From the PDB 2C1L entry:



```
CRYST1  138.925  138.925   94.135  90.00  90.00  90.00 I 4          16
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
123456789 123456789 123456789 123456789 123456789 123456789 123456789 1234567890
```

Unit cell constants
(length of the unit cell edges)
in angstroms
(Å, Ångstrøm)

Unit cell angles
in degrees

Space group
Hermann–Mauguin
symbol

Z value:
“number of
the most
populous
chains in the
cell”

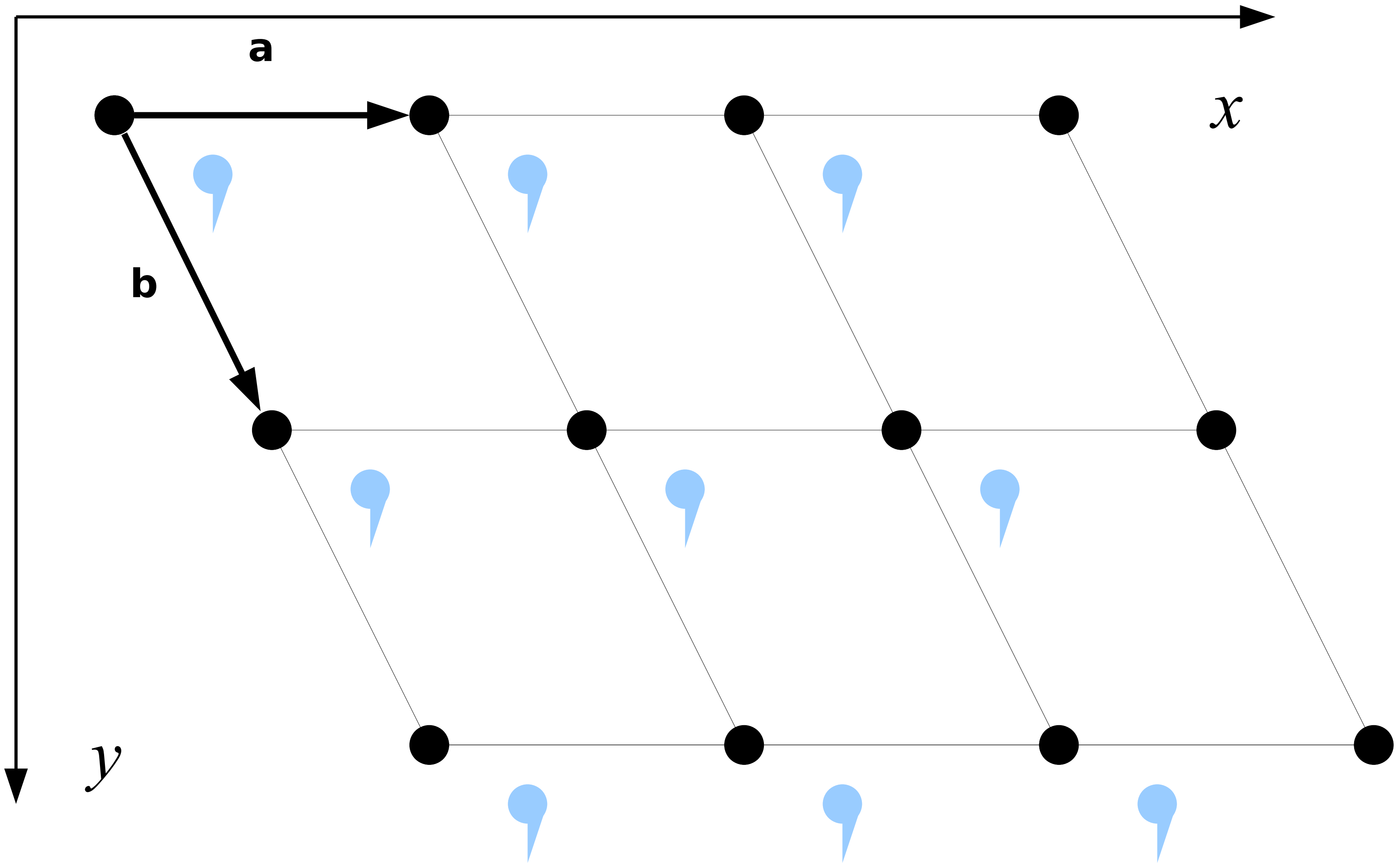
REMARK records in PDB files

- Until 1992 – free text (mostly for humans)
- From 1992 – strict layout (human and machine readable).

```
REMARK   2 RESOLUTION. 2.5 ANGSTROMS.                                   155CE  1
REMARK   3                                                              155C  15
REMARK   3 REFINEMENT. THESE ATOMIC COORDINATES MUST BE CONSIDERED AS   155CE  2
REMARK   3  PRELIMINARY. THEY WERE OBTAINED BY RUNNING SEVERAL CYCLES   155C  17
REMARK   3  OF THE DIAMOND MODEL BUILDING ROUTINE ON GUIDE POINTS FOR   155C  18
REMARK   3  ATOMS MEASURED FROM THE WIRE KENDREW MODEL. ...
...
```

```
REMARK   2 RESOLUTION. 2.17 ANGSTROMS.
...
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.17
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 24.61
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 2.000
REMARK   3   DATA CUTOFF HIGH         (ABS(F)) : 2011306.160
REMARK   3   DATA CUTOFF LOW          (ABS(F)) : 0.0000
REMARK   3   COMPLETENESS (WORKING+TEST)   (%) : 93.6
REMARK   3   NUMBER OF REFLECTIONS             : 42686
....
```

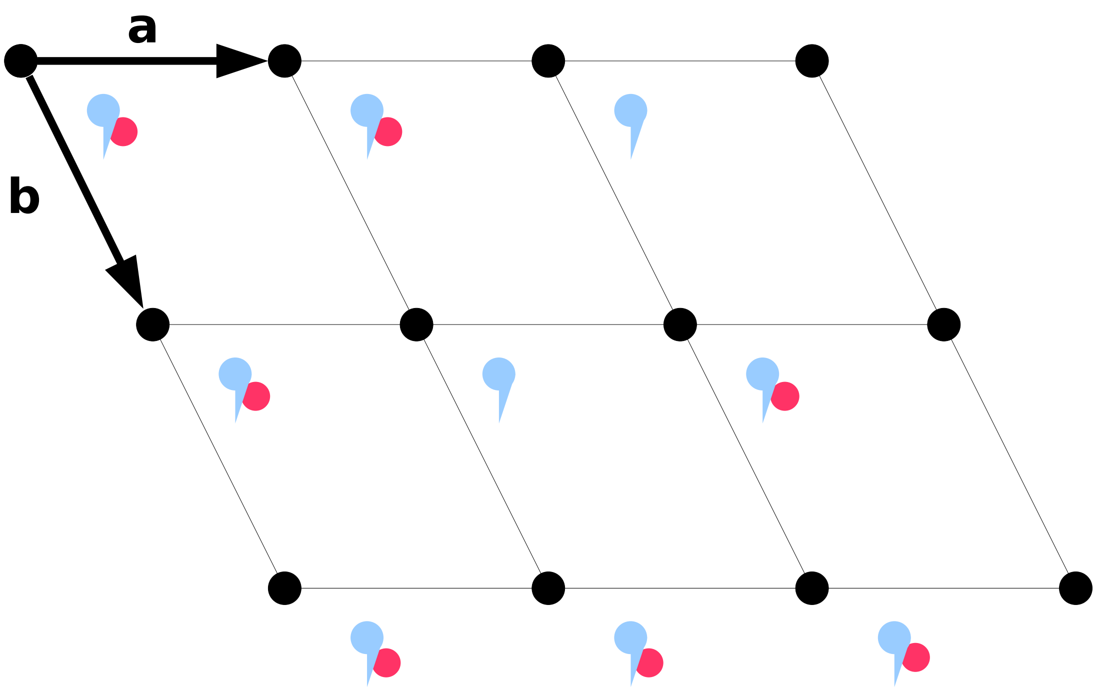
Crystal



Occupancy

- In an ideal crystal, all unit cells are identical
- In a real crystal, some atoms may be missing in some unit cells:

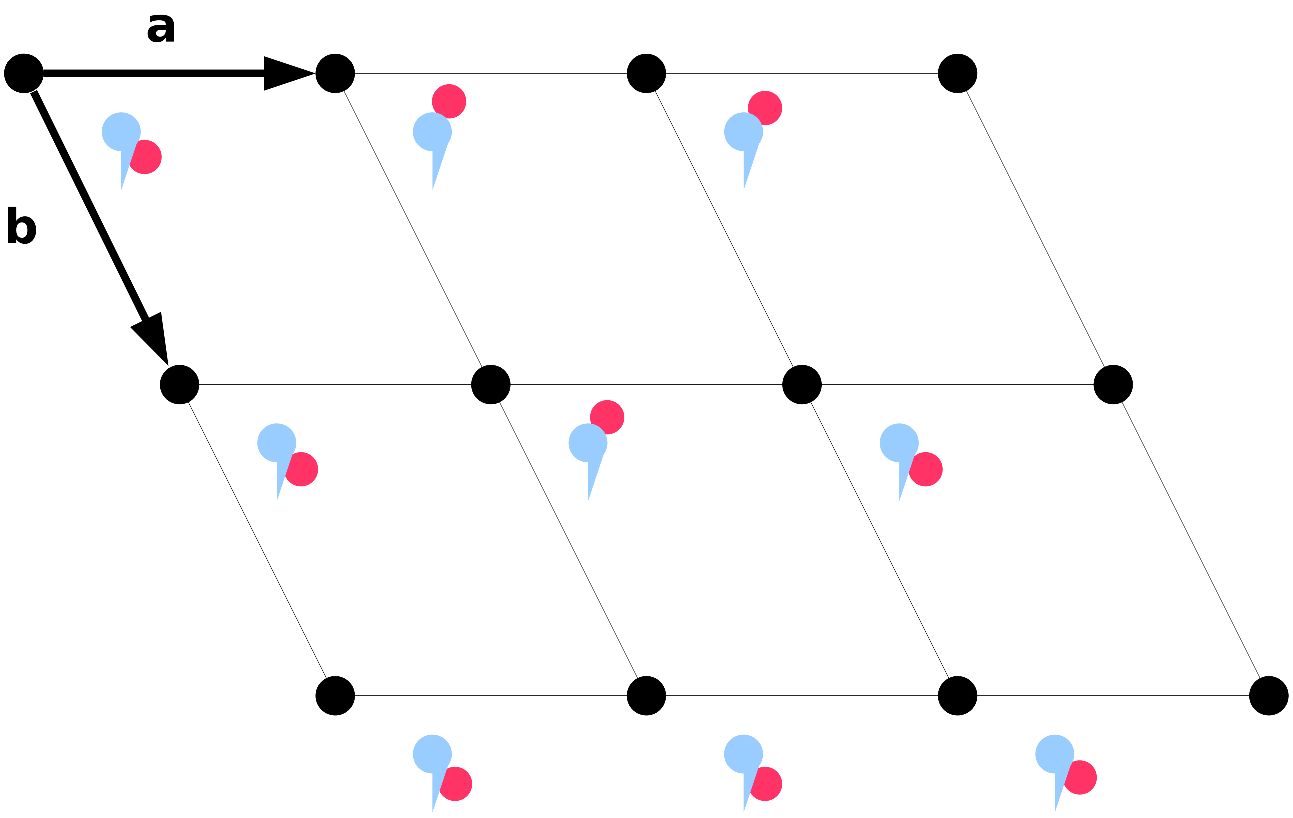
$$occupancy = q = \frac{number\ of\ unit\ cells\ with\ the\ atom}{total\ number\ of\ unit\ cells}$$

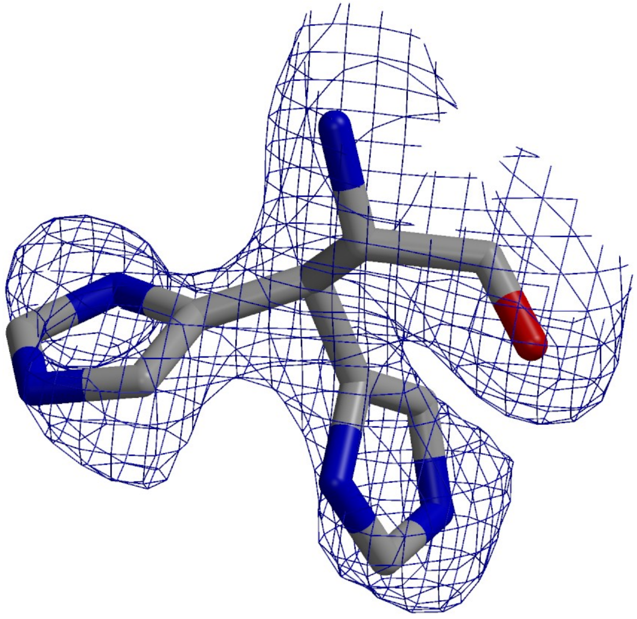


$$q = \frac{7}{9} = 0.778$$

Alternative locations

- Some atoms can have different coordinates in different unit cells:

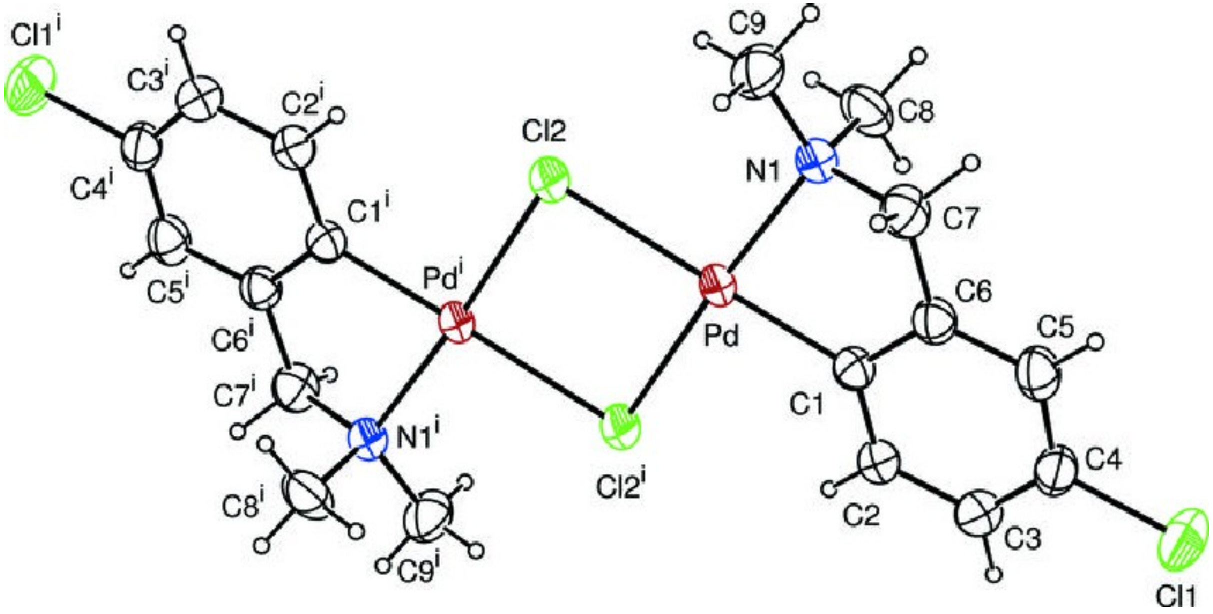




PDB ID 1KNV, His B169
Grazulis *et al.*

Temperature factors

(B-factors; IUCr: Debye-Waller factor)



Sang et al. Acta Cryst. (2010). E66, m252
http://journals.iucr.org/e/issues/2010/03/00/bq2191/index.html

$$B = 8\pi^2 \langle u^2 \rangle$$

B – Temp. factor in an ATOM record
u – RMS deviation of the atomic position.
$\langle\rangle$ – average in time

$$B = 79\text{Å}^2 \Leftrightarrow \sqrt{\langle u^2 \rangle} = 1.0\text{Å}$$

$$\langle u^2 \rangle = \lim_{T \to +\infty} \frac{1}{T} \int_{-T/2}^{T/2} u^2(t)\, dt$$

$$p(u) = \frac{1}{\sqrt{2\pi \langle u^2 \rangle}} e^{-\frac{u^2}{2\langle u^2 \rangle}}$$

Explanation of B-factors:
http://spdbv.vital-it.ch/TheMolecularLevel/ModQual/#Temperature%20factor%20(crystallogr
http://pldserver1.biochem.queensu.ca/~rlc/work/teaching/definitions.shtml
C. Giacovazzo *et al.* Fundamentals of Crystallography, IUCr & Oxford Uni. Press, p. 149

Advantages of the PDB format

- ASCII (ANSI X3.4-1986) text, human and machine readable,
- Simple
- Relatively easy to process with any program (grep, Perl, awk, C++, Python, Java)
- Widespread, well documented and standardised

Drawbacks of the PDB format

- **Fixed columns**
- No means to include X-ray, BMR or EM original data
- Awkward when additional data items need to be included
- No “official” support for multilingual text or scientific notation
  (i.e. no Unicode/UTF-8 support)