Bioinformatika III

Trimačių struktūrų analizė ir spėjimas

Paskaita 2 – struktūrinių failų formatai (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.

Struktūriniai duomenys

- Koordinatės
- Temperatūriniai faktoriai, užimtumai
- Kristalografinė informacija (gardelės parametrai, kristalo simetrija)
- Duomenų tikslumas, struktūros kokybė
- Papildoma informacija:
  - seka
  - antrinės struktūros priskyrimas
  - biocheminiai, biologiniai duomenys ...

PDB failų formatas

- ASCII koduotės tekstiniai failai
- Fiksuotų kolonėlių formatas
- Įrašas – viena eilutė
- Kiekvienas įrašas prasideda raktiniu žodžiu

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

PDB failo pavyzdys

```
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 formato ATOM įrašai

Iš originalių PDB 1KNV bei 2EZV įrašų:

```
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
```

raktinis žodis

atomo tipas, numeris, liekana – unikalus atomo identifikatorius

ortogonalios koordinatės, Å

užimtumas ir B-faktorius

atomo cheminis simbolis

```
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
```

alternatyvios padėties žymė

segmento vardas

```
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
```

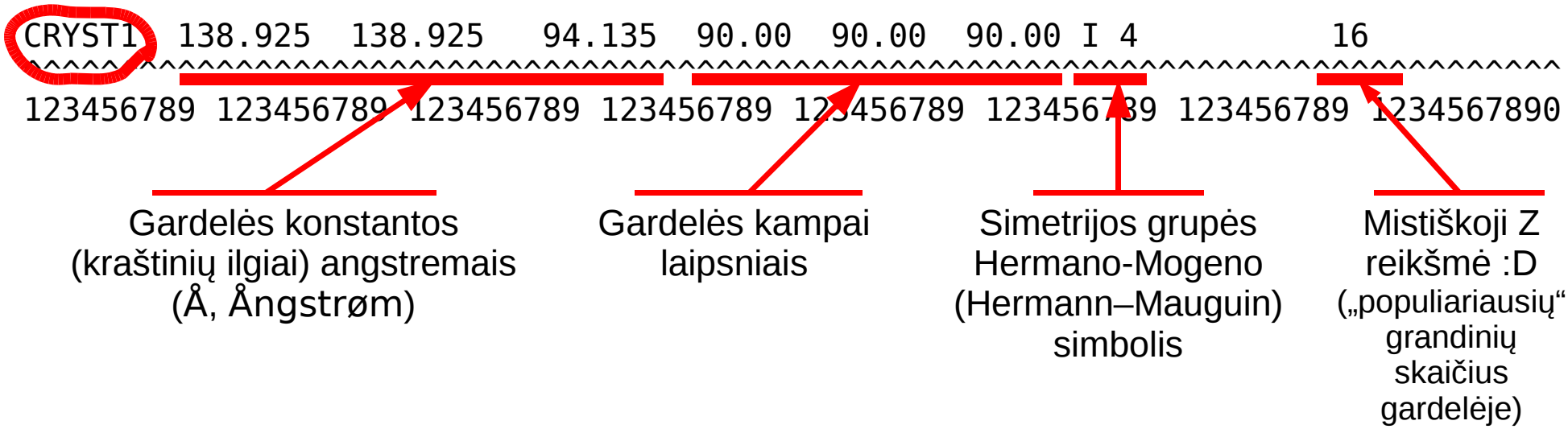
cheminis simbolis seniau buvo žymimas vado pozicijos parinkimu

atomo cheminis simbolis

krūvis

Kristalografinė informacija PDB faile – CRYST1 įrašas

Iš PDB 2C1L:



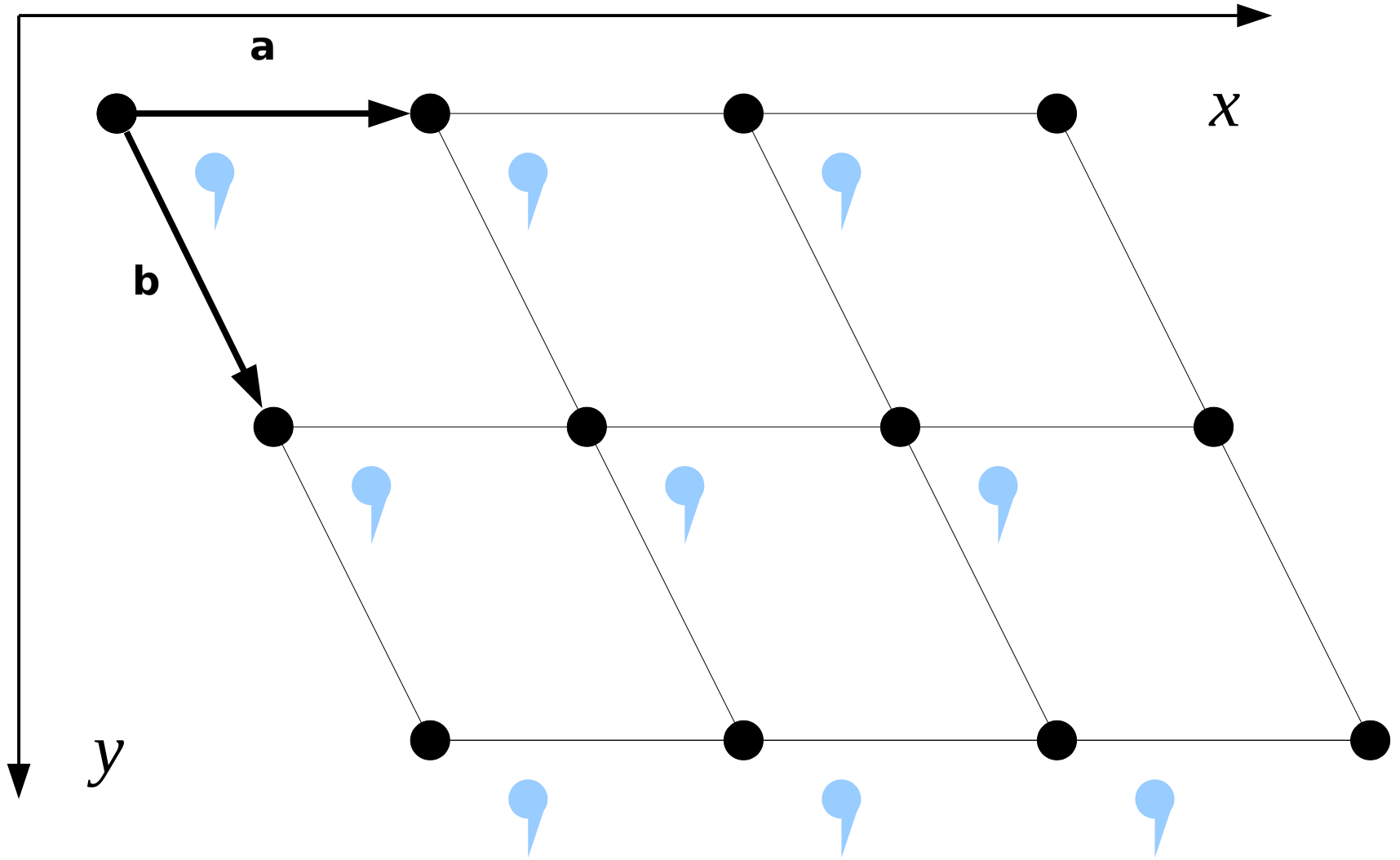
REMARK įrašai PDB failuose

- Iki 1992 – laisvas tekstas (tik žmogui)
- Nuo 1992 – griežtesnė struktūra (žmogui ir mašinai).

```
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. ...
...
```

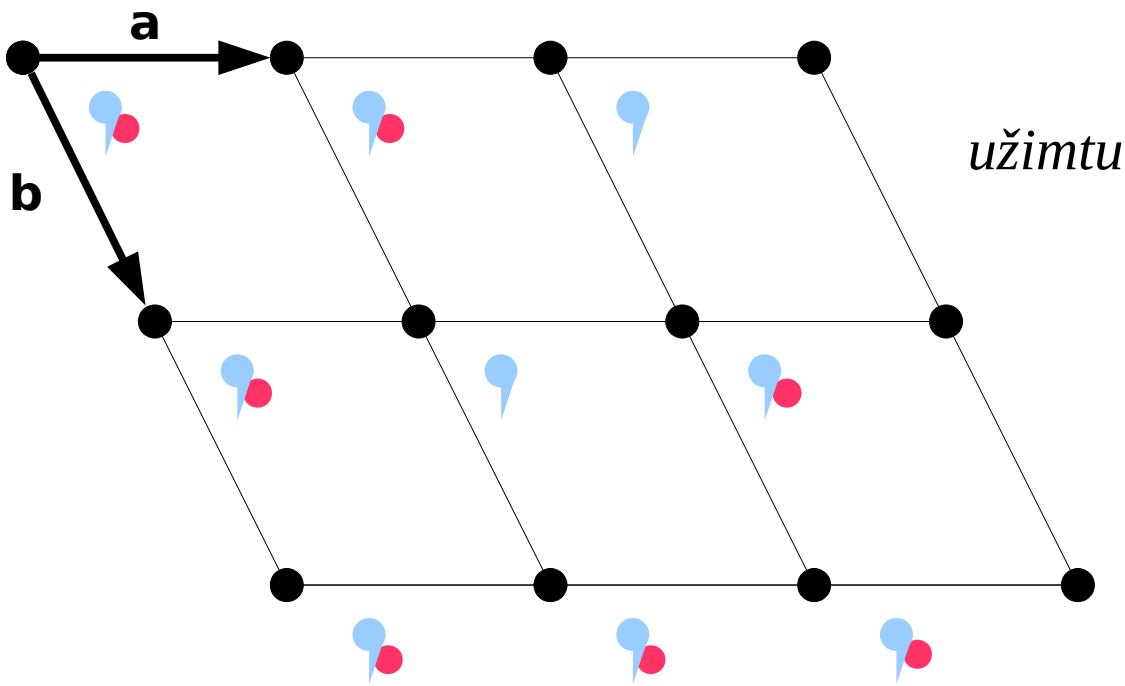
```
REMARK   2 RESOLUTION. 2.17 ANGSTROMS.
...
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.17
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 24.61
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 2.000
REMARK   3   DATA CUTOFF HIGH         (ABS(F)) : 2011306.160
REMARK   3   DATA CUTOFF LOW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

Kristalai



Užimtumas
(*angl.* occupancy)

- Idealiame kristale visų gardelių turinys yra vienodas
- Realiame kristale kai kurie atomai gali būti ne visose gardelėse:

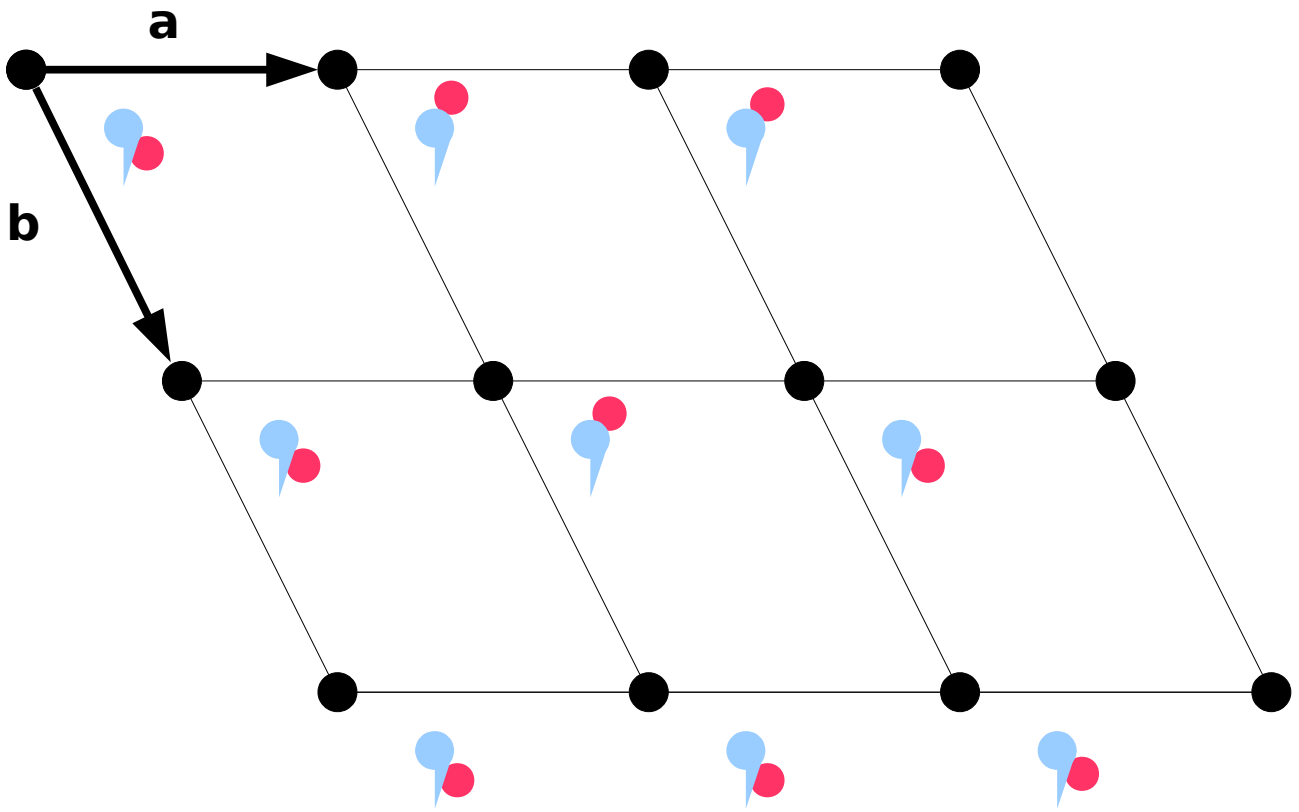


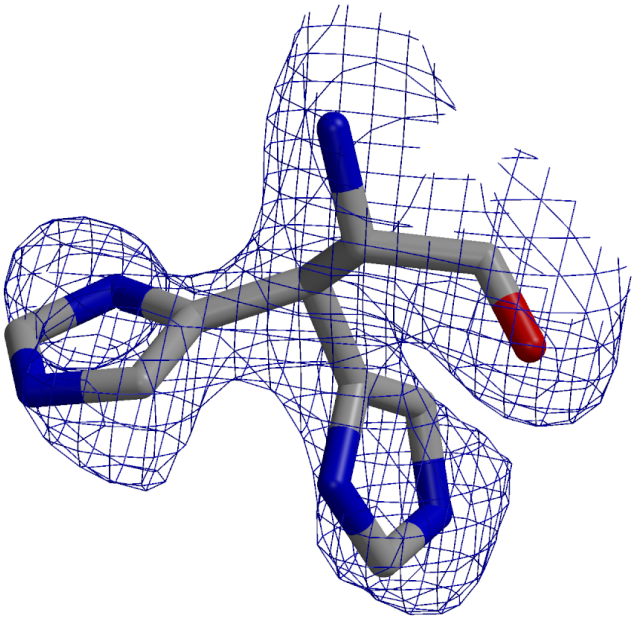
$$užimtumas = q = \frac{užimtų\ vietų\ skaičius}{bendras\ vietų\ skaičius}$$

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

Alternatyvios padėtys
(*angl.* alternative locations)

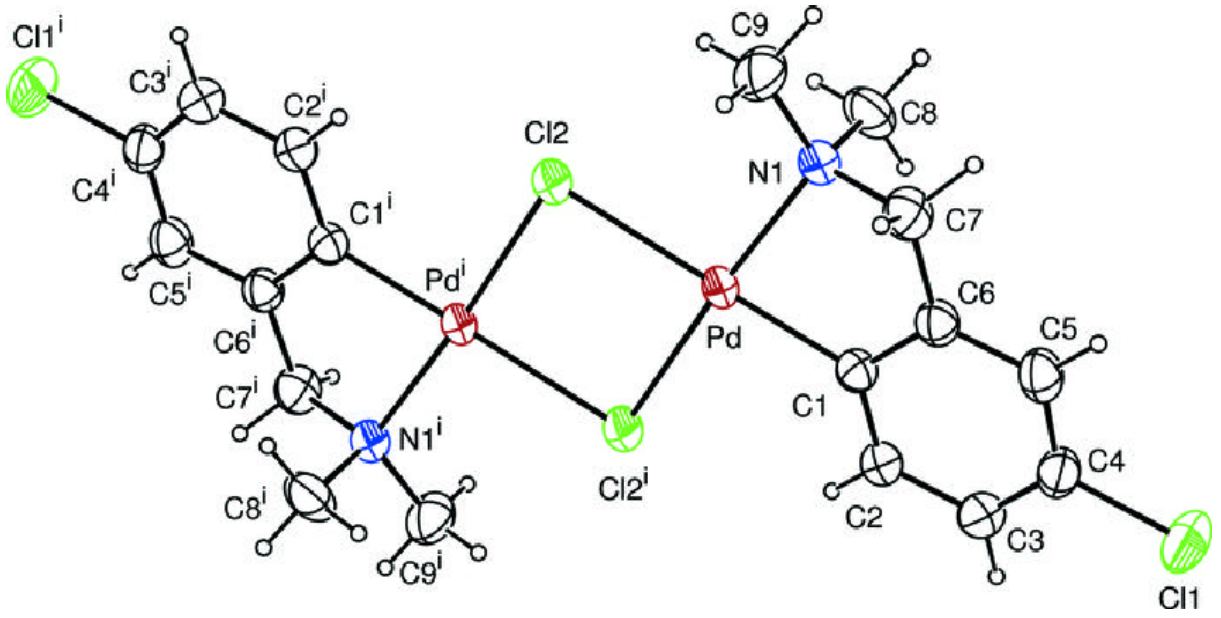
- Kai kurie atomai skirtingose gardelėse gali būti išsidėstę skirtingai:





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

„Temperatūriniai faktoriai“ (*angl.* “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 – temperatūrinis faktorius ATOM įraše
u – atomo pozicijos nuokrypis nuo vidutinės pozicijos.
⟨⟩ – vidurkis laike

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

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

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

B-faktorių paaiškinimas:
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

PDB formato privalumai

- ASCII (ANSI X3.4-1986) tekstas, įskaitomas tiek žmogui, tiek mašinai
- Formatas paprastas
- Lengva analizuoti failus nesudėtingais metodais (grep, Perl, awk)
- Paplitęs, labai gerai dokumentuotas ir standartizuotas formatas

PDB formato trūkumai

- **Fiksuotos kolonėlės**
- Nėra galimybės įtraukti į failą difrakcijos, BMR ar EM pradinius duomenis
- Kiek kėbloka įtraukti papildomus duomenų laukus
- Nėra „oficialios“ galimybės naudoti internacionalinius rašmenis bei mokslininkų priimtą notaciją (t.y. trūksta Unikodo/UTF-8 palaikymo)