Software status

and InChI version 2

InChI Software releases

- 1.00 Apr 2005

  The beginning

- 1.01 Aug 2006

  InChI2Struct and many other things appear

- 1.02 beta Sep 2007

  Introduced InChIKey (experimental), API changes

- 1.02 final Jan 2009

  Introduced Standard InChI
  (only Standard supported; InChIKey layout changed)

- 1.03 Jun 2010

  Both Standard and Non-std InChI/Key now supported

InChI Software releases

- 1.04 Sep 2011

  Maintenance release;
  more permissive license

- 1.05

  Introduced (experimental) support of polymers,
  large molecules, V3000 Molfiles;
  novel API section; multi-threading

  pre-release Oct 2016
  update Jan 2017
  final Feb 2017

- InChI for Reactions Mar 2017

InChI Software v. 1.05 release

- Maintenance release with a number of significant new features
- All things not included there will most likely be postponed to InChI version 2

Added more elements

- Updated software to current IUPAC confirmed elements list
  - (up to 118 oganesson which closes the Period 7)

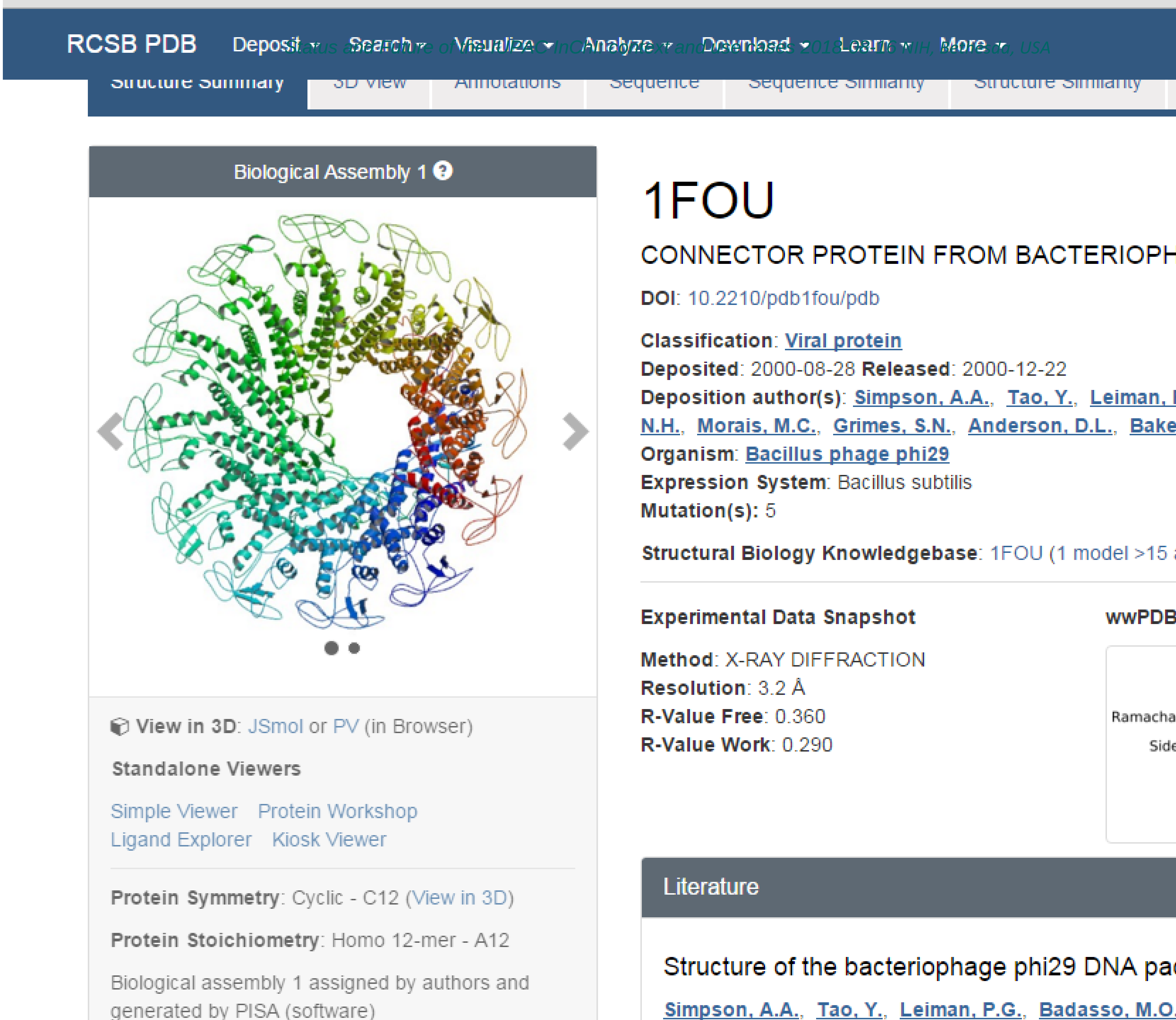
- not too much work
- but has to be done in IUPAC-endorsed software

Support of Molfile V3000 format

- Allows one to deal with
  - large (bio)molecules beyond 1000 atoms limit
  - enhanced stereochemistry (e.g., mix of Rel/Abs)
  - extended support of organometallics (haptic bonds)
- The last two features are implemented in reader but are awaiting a future use in InChI 2
- Large molecules (> 1000 atoms) reading

Support of large molecules

- Limit of number of atoms increased (technically) from 1023 to 32767
- Some other internal limits relaxed
- May be extended further
- Tests on PDB
  100,000+ molecules;
  PDB –(OpenBabel) MOL V3000 → mol2inchi



1FOU

CONNECTOR PROTEIN FROM BACTERIOPH

DOI: 10.2210/pdb1fou/pdb

Classification: Viral protein

Deposited: 2000-08-28 **Released**: 2000-12-22

Deposition author(s): Simpson, A.A., Tao, Y., Leiman, I N.H., Morais, M.C., Grimes, S.N., Anderson, D.L., Bake

Organism: Bacillus phage phi29

Expression System: Bacillus subtilis

Mutation(s): 5

Structural Biology Knowledgebase: 1FOU (1 model >15

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 3.2 Å

R-Value Free: 0.360

R-Value Work: 0.290

View in 3D: JSmol or PV (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer Kiosk Viewer

Protein Symmetry: Cyclic - C12 (View in 3D)

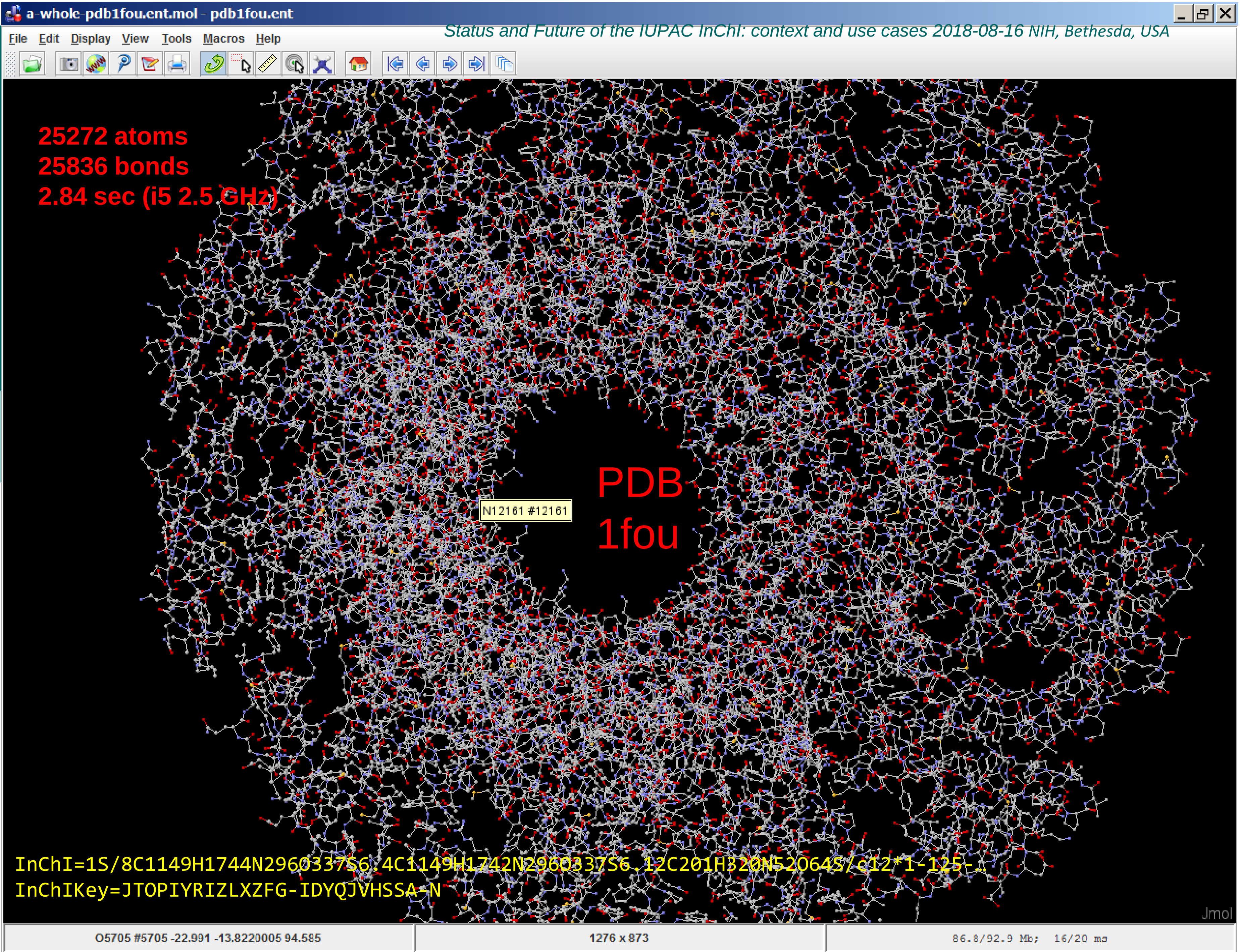
Protein Stoichiometry: Homo 12-mer - A12

Biological assembly 1 assigned by authors and generated by PISA (software)

Literature

Structure of the bacteriophage phi29 DNA pac

Simpson, A.A., Tao, Y., Leiman, P.G., Badasso, M.O.



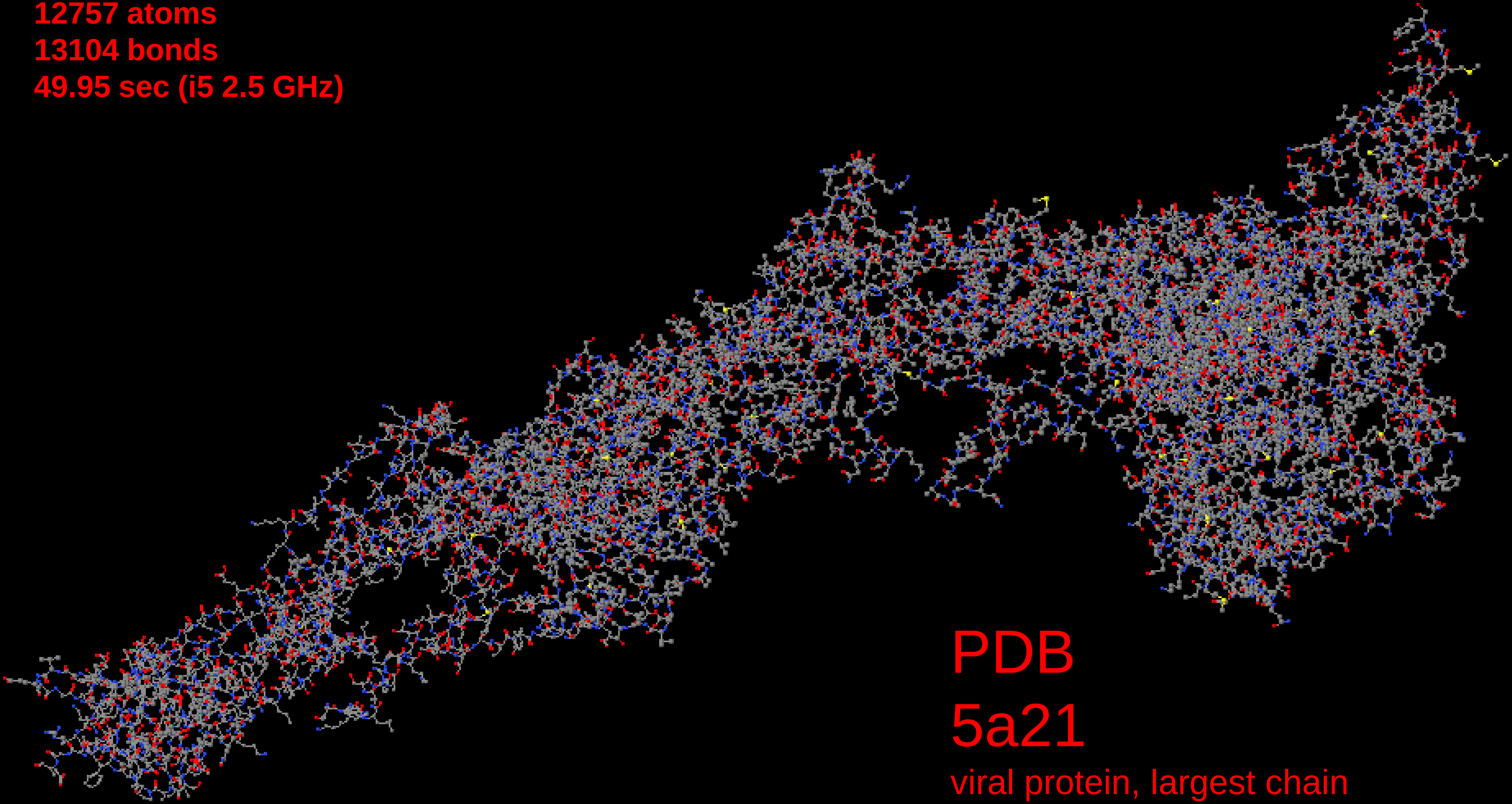
25272 atoms
25836 bonds
2.84 sec (i5 2.5 GHz)

PDB
1fou

InChI=1S/8C1149H1744N296O337S6.4C1149H1742N296O337S6.12C201H320N52O64S/c12*1-125-…
InChIKey=JTOPIYRIZLXZFG-IDYQJVHSSA-N

O5705 #5705 -22.991 -13.8220005 94.585 | 1276 x 873 | 86.8/92.9 Mb; 16/20 ms

12757 atoms
13104 bonds
49.95 sec (i5 2.5 GHz)

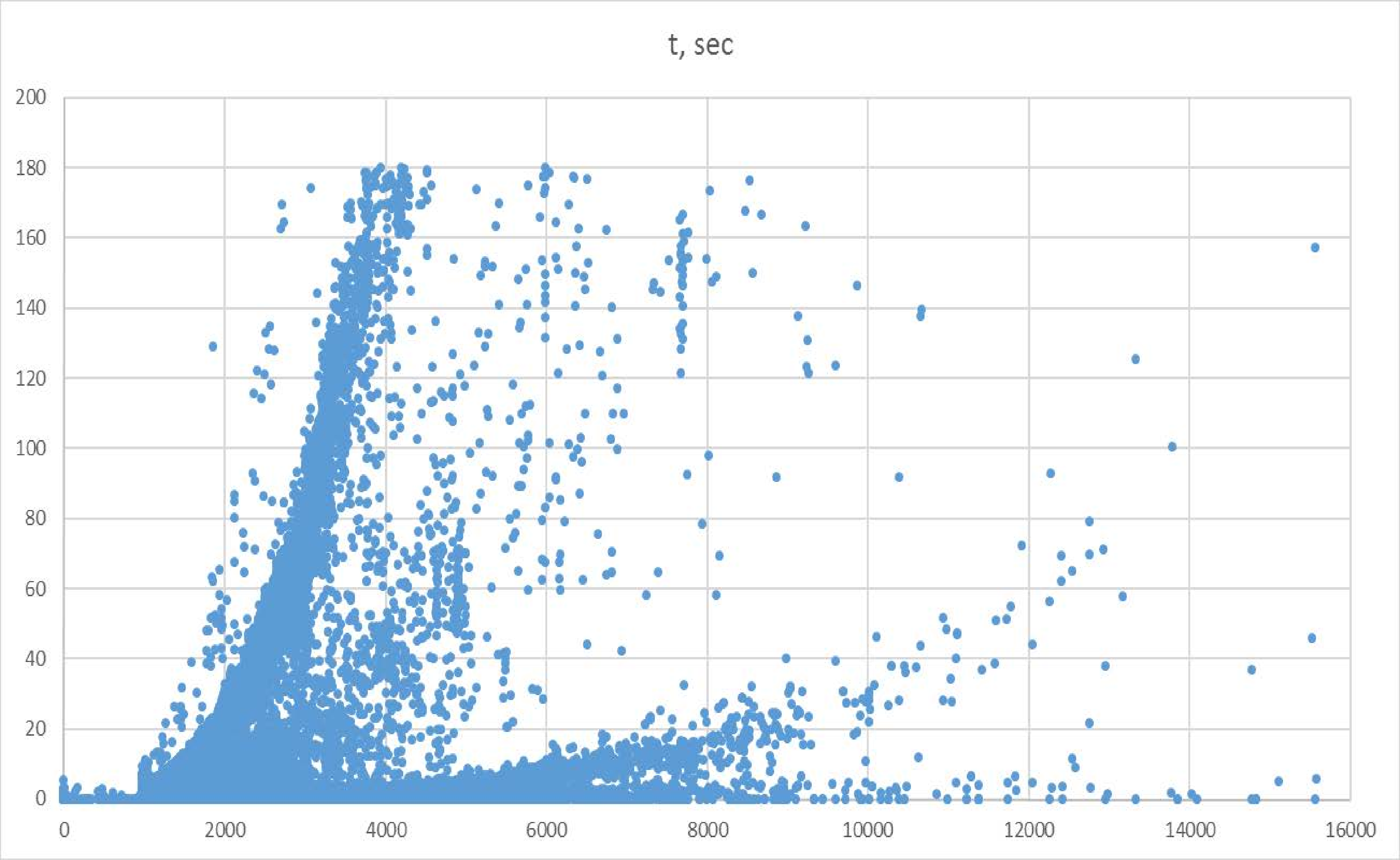


PDB
5a21
viral protein, largest chain

InChI=1S/C8080H12360N2105O2526S46/c1-852-952-1654-4476-2684-2735-8046(4476)7957(12634)9769…
InChIKey=SBVFWJWLGWCUFW-BDSVIIDHSA-N

Cautionary notes

- Is speed a concern?
- Benchmarks: i5 2.5 GHz CPU (single-core) SSD
- 99% of longest chains of ~100,000 proteins of PDB (up to ~16,000 atoms) converted to InChI for <=180 sec
- Average processing time ~ 3.8 sec (average size 2400 atoms)
- Still, there are molecules not converted to InChI for reasonable time...


t, sec
200
180
160
140
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000
14000
16000

Cautionary notes

- InChI was not designed with >> 1000 atoms in mind
- Though canonicalization and normalizations algorithms principally should work...
- and no problems were reported yet...
- several issues were already found by internal tests

Cautionary notes

- Renumbering tests
- ~70,000 max-length protein chains from PDB were tested, with 100 random atomic renumberings for each
- 14 failures detected so far
  - that is, 14 molecules from PDB give different InChI/Key’s on re-numberings
- No final clarity yet
  - problem may lie in normalization (mobile H) rather then in canonicalization

Cautionary notes

- InChI’s are getting very long
- InChIKey in its current form may be too short to serve for all the large molecules people may start to play with
- Experimental (beta) large-mol InChI/Keys are isolated from others by using ‘B’

Support of polymers

- Only simple polymers (no cross-linked, etc.)
- Source-based representation
- Structure-based representation

Known issues with polymers

- Issues on elucidation of canonical SRU
  - reported by Roger Sayle and John Mayfield, re-iterated today
  - BTW: explicitly stated in documentation (in part)
- Issue #1, simplified:
    - -[-CH2CH2-]n- *NE* -[-CH2-]n-
    - But should it? Odd/even, etc., repeatability
- Issue #2:
  - no polymer SRU “frame shift” analyzed when explicit end groups specified
    H2N-[-CH2-C(O)-NH-]n-CH2-C(O)OH *NE* H2N-CH2-[-C(O)-NH-CH2-]n-C(O)OH *NE* H2N-CH2-C(O)-[-NH-CH2-C(O)-]n-OH
  - BTW: frame shift is of course supported when star atoms (*) are shown instead
    -[-CH2-C(O)-NH-]n- *EQ* *-[-C(O)-NH-CH2-]n-* *EQ* *-[-NH-CH2-C(O)-]n-*
- In principle, solvable
  - But solution seems to be far from nice
  - Further feedback desired (this meeting, discussions, opinions of polymer chemists?)

New "extensible" (IXA) API

- IXA stands for “InChI Extensible API”
- Adds new API procedures including low-level functions to deal with atoms, bonds, etc.
- Code supplied by Digital Chemistry

  John Barnard with co-workers
- Ported to Linux
  with help of Pubchem team

  Evan Bolton, Paul Thiessen
- No problems reported (yet)

Support of safe multi-thread execution

- Allows one to significantly increase speed of InChI/Key generation while calling InChI Library on multi-CPU hardware (most of modern systems)

- Code changes supplied by Bio-Rad
  - Karl Nedwed

- Porting/testing on Linux with help of Pubchem team

  Evan Bolton, Paul Thiessen

- No problems reported (yet)

Current status

- To early to remove “experimental” label from both large molecules and polymers
- 1.051 intermediate release
  - to include fixes for several already found minor bugs & ”features”
  - may be launched on Fall 2017

Suggested near future updates

- 1.051 intermediate release
  - to include fixes for several minor bugs & ”features”
  - tentatively planned for Fall 2017

InChI version 2

- Working groups

  …

InChI version 2

- Very rough estimate (0-5) of implementation effort
  - Tautomerism 3.5
    - moderate to significant
  - Organometallics 4.5
    - significant to monstrous
  - Advanced large molecules 4.5
    - significant to monstrous

      (depend on canonicalization issues, HELM integration, ...)

InChI version 2

- Very rough estimate of implementation effort
  - QR-codes 2
    - minor
  - Mixtures 2.5
    - minor to moderate

InChI version 2

- Other (no working-groups)
- Enhanced stereo (following V3000)
  - Collections, ABS/AND/OR
    - Relatively straightforward

InChI version 2

- Other (no working-groups)

- Longer InChIKey
  - “codebreaking” sport
  - anyway, 1st block is not a real issue
    - Tolerate ~1*10^9 entries
    - (Andrey Erin: 12 collisions per 27*10^9, theor. estimate is ~10)
    - may be slightly increased in length

  - 2nd block is what really counts!

InChI version 2

- Longer InChIKey

2nd block may really have problems

There are much things there already (think of carbohydrates!) ...
people are trying to squeeze everything in there (polymers...mixtures...)
and this likely will continue)

- Make 2nd block significantly longer
- - or just add 3rd car to the train?