Stereochemical statistics in Crystallography Open Database

Andrius Merkys[a], Fei Long[b], Garib N. Murshudov[b] and Saulius Gražulis[a]

[a]Department of Protein-DNA Interactions, Vilnius University Institute of Biotechnology, V.A. Graiciuno 8, LT-02241 Vilnius, Lithuania; [b]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, England.

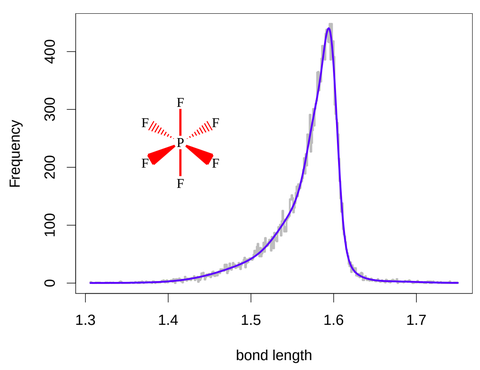
Abstract

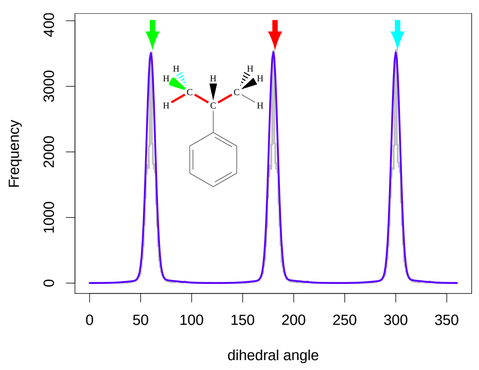
Libraries of small molecule stereochemical information, used for refinement and validation of small molecules and macromolecule-ligand complexes, are subjected to two limitations: licensing and possibility to become outdated [5]. A novel library of small molecule stereochemical information is constructed from the Crystallography Open Database [4], harnessing a new method for description of the variety of small molecule chemical environments and Bayesian framework. Means for automatic renewal of the library in the real time are devised. The result of the research is comparable to the previous works [1, 2] and proves to be useful in the detection of unusual geometric features in small molecules.

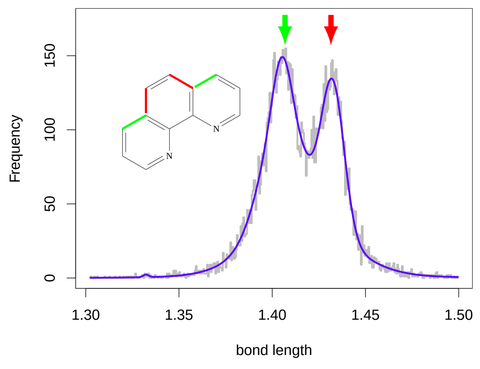
Materials and methods

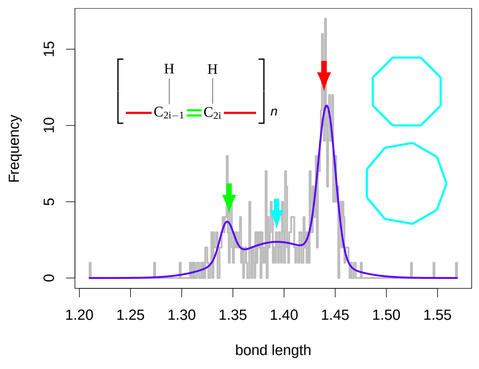
- Data source: Crystallography Open Database (COD)
- Stereochemistry extracted from ∼160 K structures:
  - ∼11 M bond lengths;
  - ∼22 M valence angle sizes;
  - ∼39 M dihedral angle sizes;
- Harnessing SMILES-like algorithm to describe atom chemical environment:
  - ∼1 M classes of bonds;
  - ∼3.5 M classes of valence angles;
  - ∼8 M classes of dihedral angles;
- Fitting mixtures to distributions:
  - Using Gaussian and Cauchy mixtures for bond lengths and valence angles;
  - Using von Mises mixtures for dihedral angles;
- Estimating mixture parameters with expectation-maximisation algorithm [3];
- Selecting best models using Bayesian information criterion [6].

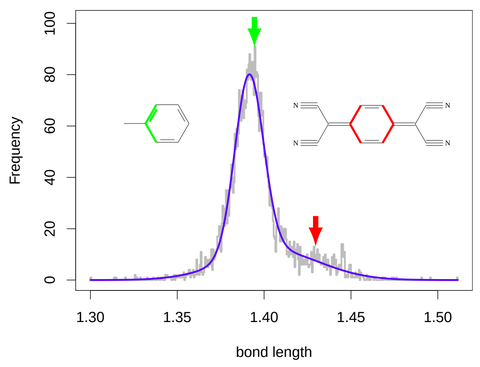
Fitting the distributions

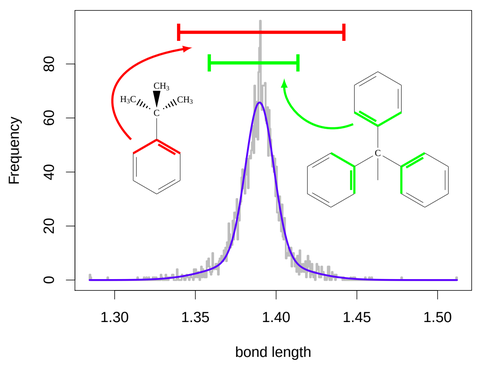












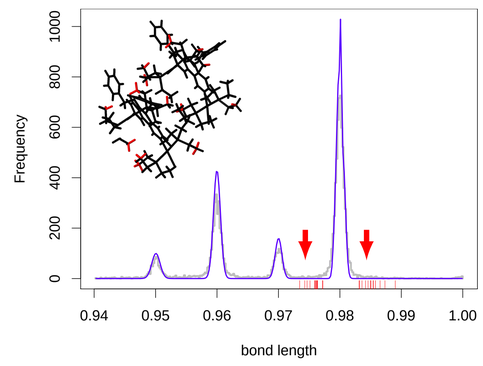
- Complex distributions can be described (*fig. a,b*);
- Information is preserved even when classification algorithm is not accurate enough (*fig. c,d,e,f, coloured bars represent* $\pm 3\sigma$).

a	b
c	d
e	f

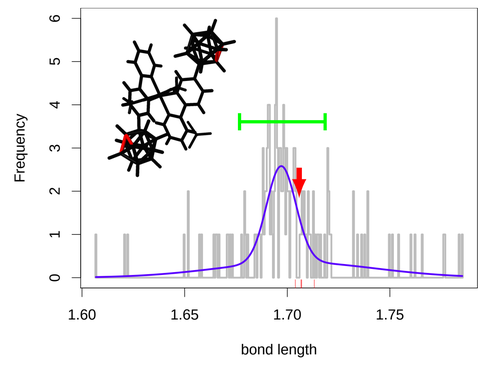
Detecting unusual features in molecular models

- Depends on a set of models defining ”normal” geometry;
- ”Unusual” does not necessarily mean ”wrong”.

- COD structure 4027109
  - Methyl C-H bond lengths deviate from ”optima” (*up right*);
  - Do deviations notify about structure anomalies or represent previously unobserved albeit normal bond lengths?

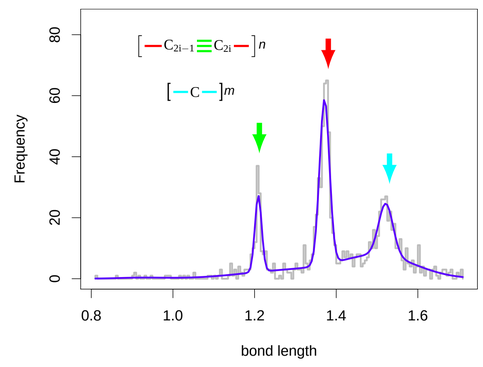


- COD structure 7109296
  - Carborane C-B bonds fall in ”credible” interval (*down right, green bar marks* $\pm 3\sigma$ *from the centre of the sharpest peak*);
  - Formally, carborane carbon is hexavalent, though such structures are reliably observed.

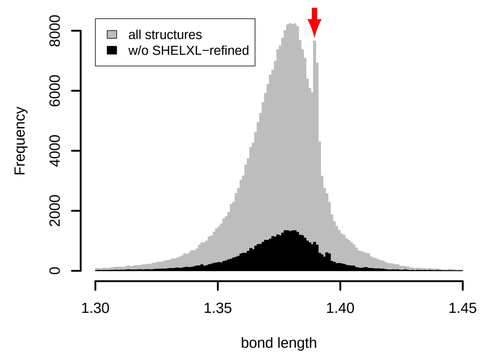


Insights in aggregated published material

- Problems with model representation introduce non-physical modes:
  - Example: cyan peak in linear carbon bond histogram (*up right*) marks bonds from alkanes without covalent-bounded hydrogens;
  - Such models should undergo special treatment or be removed from dataset;
  - Requires attention of an expert to be detected.



- Refinement constraints bias the distributions
  - Example: red-marked peak in benzene C-C bonds histogram (*down right*);
  - Removing observations of constrained models suppresses bias.



Conclusions

- Fully automatic and unsupervised methods to collect and process the stereochemical parameters are devised;
- Over 160 K small molecule structures were used to build stereochemical database;
- Database proved to be useful for the detection of unusual geometric features in molecular models;
- May be applied for maximum likelihood refinement of structures of macromolecule/ligand complexes.

Acknowledgements

This research was funded by a grant (No. MIP-025/2013) from the Research Council of Lithuania.

Bibliography

[1] Allen et al. Tables of bond lengths determined by X-ray and neutron diffraction. part 1. bond lengths in organic compounds. *J. Chem. Soc., Perkin Trans. 2*, pages S1–S19, 1987.

[2] Andrejašič et al. PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. *Acta Crystallographica Section D, Biological Crystallography*, 64(Pt 11):1093–109, 2008.

[3] Dempster et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[4] Gražulis et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427, Jan 2012.

[5] Jaskolski et al. Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta crystallographica. Section D, Biological crystallography*, 63(Pt 5):611–20, 2007.

[6] Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

andrius.merkys@gmail.com