[MS10-P14] 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, VU Institute of Biotechnology, V.A. Graiciuno 8 LT-02241 Vilnius, Lithuania;*
[b]*Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, England.*
E-mail: andrius.merkys@gmail.com

Libraries of molecular stereochemical information are subjected to two limitations: licensing and threat of becoming outdated [1]. Open and regularly updated organized collections of molecules help overcoming aforementioned limitations. A novel library of small molecule bond lengths, valence and dihedral angles is constructed from the Crystallography Open Database (COD) – an open crowdsource-powered resource of small molecule structures [2] – harnessing a new method for description of the variety of small molecule chemical environments and the Bayesian framework implemented in open source software. Gaussian, Cauchy and von Mises mixture models are used to describe the distributions of bond lengths and angle sizes for the first time in this field, outperforming the standard symmetric unimodal models. The expectation maximization algorithm [3] is employed to determine model parameters, the principle of Occam's razor is implemented for model selection using Bayesian information criterion [4]. The means for automatic unsupervised renewal of the library in the real time are devised. In total the library currently contains 11 M bonds, 21 M valence angles and 39 M dihedral angles extracted from more than 150 thous. small molecule structures from COD and broken down to respectively 1 M, 4 M and 8 M classes. The result of the research is comparable to the previous works of Allen [5] and Andrejašič [6] and proves to be useful in the detection of unusual geometric features in molecular models. Mixture models and expectation maximization algorithm displays the ability to cover the whole variety of multimodal, skewed and periodic distributions of bonds and angles, originating from the structures in COD. Unsupervised methods for extraction, classification and description of small molecule stereochemical information together with constantly expanding resource of small molecule structures allow the construction of open and regularly updated knowledge database to be used in the fields of crystallographic refinement, molecule model validation, rational drug design and the research of material properties. Reported methods can be improved by devising better means of initialization of the expectation maximization algorithm and speeding up the convergence.

[1] Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007). 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, pp. 611-20.

[2] Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N.R., Moeck, P., Downs, R.T., and Le Bail, A. (2012). Crystallography Open Database (COD): an openaccess collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Research 40, pp. D420-D427.

[3] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society 39, pp. 1-38.

[4] Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, pp. 461-464.

[5] Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, A.G., and Taylor, R. (1987).

Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc., Perkin Trans. 2, pp. S1-S19.

[6] Andrejašič, M., Pražnikar, J., and Turk, D. (2008). PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures.. Acta Crystallographica Section D, Biological Crystallography 64, pp. 1093-109.

Keywords: crystallographic refinement; crystal structure databases; statistical analysis