Chemical annotation in the Crystallography Open Database

A. Merkys[1], A. Vaitkus[1], A. Grybauskas[1], A. Konovalovas[1], M. Quirós Olozábal[2], S. Gražulis[1]

[1]Vilnius University Life Sciences Center, Saulėtekio al. 7, 10257 Vilnius, Lithuania,
[2]Departamento de Química Inorgánica, Universidad de Granada, 18071, Granada, Spain

andrius.merkys@gmc.vu.lt

Reliable knowledge about structure and properties of chemical compounds is essential for pharmacology, food safety, environment preservation, design of new materials and understanding of functions of small molecules in living organisms. The number of unique substances known to humanity currently exceeds 100 million [1], and only the use of computers makes coping with the amount of available information possible.

The most accurate data about the structure of molecules are obtained from X-ray crystallographic (XRC) analyses. Currently, about 1 million crystal structures are known and the use of this information is enabled by crystallographic databases [2-4]. These data, however, are not immediately usable by chemists. XRC determines accurate 3D coordinates of each atom in a crystal, and, in extreme cases, electron densities along chemical bonds, but it does not detect atomic charges, bond types or the presence of lone electrons in radicals. All such information needs to be inferred from the crystallographic data based on current chemical knowledge, either manually [5], or using heuristics, implemented as computer programs [6-7]. However, the existing programs rarely consider information other than the coordinates. What is more, heuristics are usually specifically tailored for organic molecules. As a result, the derivation of chemical annotations by these programs is not always reliable, especially for metal-organic complexes.

Nevertheless, atomic coordinates in crystal structure reports are usually accompanied by additional chemical information. Systematic chemical names are often provided or derivable from publication titles or texts. Connectivity details in machine-readable formats may follow as well, albeit usually in forms not suitable for automated overlaying on the coordinate data. All this information could be employed to annotate crystallographic data with chemical details provided the mapping between different representations is known.

The largest open access crystallographic database, the Crystallography Open Database (COD, [2]), contains computer readable chemical descriptions for nearly half of its entries [5]. Currently, these descriptions are not linked to particular atoms in crystals, thus studies that require the combined crystallographic and chemical information have to infer the correspondence on their own. This task is tedious, involves repetition of work, and disregards readily available high-quality chemical descriptions.

Graph-based algorithms can be used to determine the links between the crystallographic and chemical data in the COD. Establishment of isomorphism between graphs derived from atomic coordinates and graphs derived from chemical descriptions enables the assignment of chemical attributes to individual bonds and atoms. Open-access nature of the COD allows dissemination of this information under FAIR (Findability, Accesibility, Interoperability and Reusability [8]) principles on the Web, immediately enabling numerous computational searches and research by pharmaceutical companies and academic groups. Thus, publishing and maintaining chemical annotations for crystallographic data in the COD would enhance research capabilities in pharmaceutical science, bio- and cheminformatics, materials science.

[1] CAS REGISTRY, https://www.cas.org/support/documentation/chemical-substances

[2] Gražulis et al. (2012). Nucleic Acids Research, 40. doi:10.1093/nar/gkr900

[3] wwPDB consortium (2019). Nucleic Acids Research, 47. doi:10.1093/nar/gky949

[4] Groom & Allen (2014). Angewandte Chemie International Edition, 53, 3. doi:10.1002/anie.201306438

[5] Quirós et al. (2018). Journal of Cheminformatics, 10, 1. doi:10.1186/s13321-018-0279-6

[6] O'Boyle et al. (2011). Journal of Cheminformatics, 3. doi:10.1186/1758-2946-3-33

[7] Willighagen et al. (2017). Journal of Cheminformatics, 9, 1. doi:10.1186/s13321-017-0220-4

[8] Wilkinson et al. (2016). Scientific Data, 3. doi:10.1038/sdata.2016.18

Keywords: Crystallography Open Database; chemistry; cheminformatics

This research has received funding from the Research Council of Lithuania under grant agreement No. MIP-20-21.