[P18] Workflow for Reaction Database Cleaning and Data Standardization

Nugmanov R.I.[1], Madzhidov T.I.[1], Varnek A.[1,2]

[1]*Kazan Federal (Volga Region) University, Butlerov Institute of Organic Chemistry, ul. Kremlevskaya 18, 420008, Kazan, Russia, e-mail:stsouko@live.ru*

[2]*Laboratory of Chemoinformatics. University of Strasbourg. 1 rue Blaise Pascal, 35000, Strasbourg, France.*

A priori assessment of optimal reaction conditions for a given transformation is the holy grail of synthetic organic chemistry. Usually, the choice of reaction conditions proceeds in essentially empirical way: the chemist relies either on his/her own experience or on information for similar reactions retrieved from the literature. However, the exponential growth of current chemical information makes the task of analysis and generalization extremely difficult for the human mind alone and requires special approaches and tools in order to efficiently extract such knowledge from raw data. Both information retrieval and automatic knowledge extraction suffer from the only one problem: despite a very large amount of reaction data exist in the literature that are manually collected by such database vendors like CAS (100 mln reactions in database) and Reaxys (60 mln reactions in database) the quality of extracted data in database is not satisfactory.

The problem could be caused not only by errors in the structure extraction but also by intrinsic features of chemical compounds like tautomerisation, epimerisation, racemization, acid-base equilibrium and others. These problems are usually solved by means of structure standardization and automatic or manual data curation procedures.

However, for databases of chemical reactions there are problems that never appeared for chemical substances additionally to regular structure standardization issues.

We present a prototype of the system for reaction standardization. It consists from different modules responsible for special steps or reaction cleaning. For the time being it could not resolve all mentioned problems but already allows solving the most common ones: atom-to-atom mapping errors detection and fixing of most common errors, unbalanced reactions completion, salts and tautomerisation representation standardization, reaction classification and its type detection. The core of the developed approach is based on CGR representation of chemical reactions.

The developed workflow returns user not only standardized reaction but also comments on possible problems with its representation, as well as the score that reflects the confidence that standardization is correct.

Bibliography:

[1] Hoonakker, F., Lachiche, N. & Varnek, A. Int. J. Artif. Intell. Tools 20, (2011) 253–270

The research was supported by Russian Scientific Foundation, grant 14-43-00024. We thank the Reaxys database (Elsevier, Netherlands) for providing us with the experimental reaction data and ChemAxon company for the software license.