[P38] Chemical reactions visualization: do outliers make any sense?

Iuri Casciuc[1], Pavel Sidorov[1,2], Alexandre Varnek[1]

[1] Laboratory of Chemoinformatics, University of Strasbourg, France
[2] Laboratory of Chemoinformatics and Molecular Modeling, Kazan Federal University, Russia

In this work we demonstrate an application of the Condensed Graph of Reaction (CGR) approach to visualization and analysis of chemical reactions databases. A CGR encodes a chemical reaction by one sole molecular graph described by both conventional bonds (single, double, etc.) and dynamical bonds (single-to-double, broken single, created double, etc.) characterizing chemical transformations[1]. Thus, a CGR can be considered as a pseudomolecule for which some types of molecular descriptors can be generated.

In this work we used dataset of 8546 reactions selected from Reflib and ChemInform databases using conventional MDL substructural queries. It consists of 6 classes of reactions: nucleophilic substitution, dihydroxylation, epoxidation, metathesis, Sonogashira and Diels-Alder. Each reaction was transferred to CGR followed by generation of ISIDA fragment descriptors[2] representing atoms and bonds sequences containing from 2 to 4 atoms. In such a way, each reaction can be considered as an object of N-dimensional descriptors space. The Generative Topographic Mapping[3] (GTM) approach has been used as a dimensionality reduction method in order to visualize the dataset in 2D latent space. GTM intrinsic parameters have been determined in a cost function optimization. Different types of cost functions have been tested: data likelihood and various measure of goodness of clustering: Γ_N function as well as distribution and distance consistency[4].

Although, almost all maps demonstrated a good ability to separate the reaction classes, some “impurities” corresponding to reactions of one class situated in more or less compact cluster of another class have been detected. Analysis of these “visualization outliers” reveals that they result from 1) errors of initial class assignment with MDL queries, 2) wrong atom-to-atom mapping, 3) rare molecular fragments, and 4) complexity of reaction mechanism (coupled reaction or multistage mechanisms). This study demonstrates that a failure of conventional MDL substructural searching queries to distinguish different classes of reactions can be easily overcome by using CGRs which provide with unique structural motifs unambiguously characterizing a given reaction class.

Bibliography:

1. Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A.,. Int. J. Artificial Intelligence Tools 2011, 20 (02), 253-270.
2. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G., Current Computer-Aided Drug Design 2008, 4 (3), 191.
3. (a) Bishop, C. M.; Svensén, M.; Williams, C. K., GTM: Neural computation 1998, 10 (1), 215-234; (b) Gaspar, H. l. n. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A., Journal of chemical information and modeling 2013, 53 (12), 3318-3325.
4. Kireeva, N. V.; Ovchinnikova, S. I.; Tetko, I. V.; Asiri, A. M.; Balakin, K. V.; Tsivadze, A. Y., ChemMedChem 2014, 9 (5), 1047-1059.