[P35] Visualization and analysis of large dataset of chemical reactions using GTM

Pavel Sidorov[1,2], Arkadii Lin[2], Timur Madzhidov[2], Alexandre Varnek[1*]

[1] *Laboratory of Chemoinformtics, University of Strasbourg, 1 rue Blaise Pascal, 67008, Strasbourg, France*
[2] *A.M. Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya st., 18, 420008, Kazan, Russia*

Analysis of chemical reactions space is still a challenge because of the complexity of reactions typically considering several molecular graphs of reactants and products. Here, we demonstrate how combination of two approaches Condensed Graph of Reaction (CGR) and Generative Topographic Mapping (GTM) could be efficiently used for this purpose.

The computations were performed on the dataset consisted of more than 48000 deprotection reactions proceeding under catalytic hydrogenation conditions. The data were curated and standardized; the atom-to-atom mapping was performed Then, each reaction has been transformed into Condensed Graph of Reaction (CGR) – 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]. Ensemble of CGRs has been encoded by ISIDA fragment descriptors[2] representing atoms and bonds sequences containing from 2 to 6 atoms. The latter served as an input into the ISIDA/GTM program, realizing Generative Topographic Mapping[3] (GTM) approach used in this work for visualization and analysis of reactions chemical space.

GTM is a non-linear dimensionality reduction, method which allows one to visualize the data in a 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 (N=5 and 20), distribution and distance consistency, kernel density estimation[4].

The “best” map based on γ_5 function provides with an interesting insight into the studied dataset. Thus, most big classes (e.g., cleavage of a benzyl group from aliphatic alcohol) form distinct clusters. The zones where several classes overlap are often populated with side reactions (e.g., double bond reduction) or with those with simultaneous cleavage of several different groups. Thus, a combination of CGR and GTM approaches looks a reasonable strategy of visualization and analysis of chemical reaction space.

Bibliography:

[1] Hoonakker F.; Lachiche N.; Varnek A.; Wagner A. J. Artificial Intelligence Tools. 20 (2011) 253.
[2] Varnek A.; Fourches D.; Hoonakker F.; Solov’ev V.P. J. Computer-Aided Molecular Design. 19 (2005) 693.
[3] Kireeva N.; Baskin I.; Gaspar H.; Horvath D.; Marcou G.; Varnek A. Mol. Inf. 31 (2012) 301.
[4] Ovchinnikova S.I. ; Bykov A.A. ; Tsivadze A.Yu. ; Dyachkov E.P. ; Kireeva N.V. J. Cheminform. 6 (2014) 20.