A. Khayrullina [1]
T. Madzhidov [1]
R. Nugmanov [1]
I. Baskin [1,2]
A. Varnek [1,3]

DEVELOPMENT OF ALGORITHM FOR CREATING ATOM-ATOMIC MAPPING USING "NAIVE" BAYES MACHINE LEARNING METHOD

[1] Chair of Organic Chemistry, A.M. Butlerov Insitute of Chemistry, KFU, Kremlevskaya St., 18. 420008 Kazan, Tatarstan Republic, Russia.
[2] Faculty of Physics, M.V.Lomonosov Moscow State University, 119991 Moscow, Russia
[3] Laboratory of Chemoinformatics, University of Strasbourg. 1 rue Blaise Pascal, 35000 Strasbourg, France.

e-mail: adelenok1993@list.ru

The chemical reaction is the conversion one or more substrate into products that differs from them in the chemical composition or structure. Knowledge of the reaction mechanism allows us to describe in detail the changes that occur at each elementary stage or for several stages.

The fundamental first step in the computer analysis of chemical reactions is determination of the correspondence between atoms of substrats and products, called the atom-atom mapping (AAM) [1]. AAM is used to find the changing part of substrate and product molecules, i.e. the reaction center [2]. Knowing reaction center it is possible to run advanced reaction search, like substructure and similarity search, establish reaction type, etc. Usually special algorythms are used to establish AAM which is significantly faster than manual curation. The most well-known and consummate algoryths are implemented in EPAM Indigo [3], Accelrys Automapper [4], JChem Standardizer [5] and ICMap [6] programs. All of them are based on maximum common substructure search, however other approaches exist as well [2]. Anyway, AAM establishing is a NP-complete problem, and thus either suboptimal solutions could be found in reasonable time or one have to use costly optimization approaches to find optimal solution.

In this work we propose a novel approach to find optimal AAM that is based on application of machine learning techniques. The task is formulated as classification: for every pair of reagent-product atom one need to establish whether this mapping is correct or not. For training classifier, pairs of atoms that correspond to correct and incorrect AAM were generated for every reaction. A simple probabilistic "Naive" Bayesian classifier (NB) was used [7]. The attribute vector for every (reagent atom, product atom) pair united information on environment of both atoms, represented by fragment descriptors. Different schemes were used to make the sole attribute vector for the atom pair based on concatenation, multiplication, etc. Thus, for a given atom pair from test set the probability that this pair correspons to correct AAM is returned. Using Muncres algoryth [8] mapping of atoms from product to reagent that correspond to maximum likelihood is identified. Special approaches were added to correctly handle molecular symmetry.

We show on some examples that the proposed algorythm is able to identify AAM and is almost as correct as other approaches. It is the first example of self-learning algorythm for AAM establishing.

1. Varnek A. et al. *J. Comput. Aided. Mol. Des.*, 2005, **19**: 693–703.
2. Chen W.L. et al. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2013, **3**: 560–593.
3. Indigo. http://epam.github.io/lifescience/indigo/index.html.
4. Moock T.E. et al. *In: Chemical Structures, The International Language of Chemistry. Berlin: Springer-Verlag*, 1988: 303-313.
5. ChemAxon JChem. JChem Base. https://www.chemaxon.com/products/jchem-base/
6. ICMAP. http://www.infochem.de/products/software/icmap.shtml.
7. Baskin I.I. et al. *Kazan Federal University*, 2016, **4**: 95-104.
8. Munkres J. *J. Soc. Indust. Appl. Math.*, 1957, **5**: 32-38.

The research was supported by Russian Scientific Foundation, grant 14-43-00024