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

MOLECULE REPRESENTATION AND DATABASING OF MOLECULES AND REACTIONS

[1] A.M. Butlerov Insitute of Chemistry, Kazan Federal University, Russia
[2] University of Strasbourg, Strasbourg, France

timur.madzhidov@kpfu.ru

The amount of data, existing in chemistry expands explosively. The recent trend shows that every 5 years the number of known compounds increases twice. Recently more than 130 mln compounds and 95 mln reactions are collected in the largest databases, such as CAS. Such a large amount of data need to be stored and effectively explored.

In the first part of the lecture molecule databasing will be reviewed. Effective storage of molecules requires compact and comprehensive computer representation, brief overview of which will be provided.

Having database of molecule one need to use the information and thus algorythms for information retrieval are of cricial importance. The following molecular search types are the most widely used: structure, substructure and similarity searches. The explanation of the core of the algorythms will be provided.

The second part of the lecture will be dedicated to the reaction storage in databases and the retrieval of required information. The most common representations of reactions in databases will be reviewed. Special attention will be paid to the problem of the atom-to-atom establishing that is cricial for certain tytpes of reaction storage and search. Common reaction search types include search over reactions and molecules involved in the transformation. The latter are very similar to usual molecular search and uses a very similar algorythms. The former include reaction structure, substructure and similarity searches. Common approaches that implement corresponding reaction search type will be reviewed.

And finally we will show how Condensed Graph of Reaction could help to implement reaction and molecular search in a unified way.

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