O. Tarasova[1],
A. Urusova[1],
A. Zakharov[2],
M. Nicklaus[2],
V. Poroikov[1]

APPLICATION OF THE LARGE-SCALE DATABASE TO THE QSAR MODELING OF THE HIV-1 REVERSE TRANSCRIPTASE INHIBITORS

[1] Laboratory for Structure-Function Based Drug Design, Institute of Biomedical Chemistry, Pogodinskaya Str., 10 Building 8, Moscow, Russia, 119121;
[2] CADD Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, DHHS, NCI-Frederick, Building 376, Room 205, 376 Boyles St., Frederick, MD 21702

olga.a.tarasova@gmail.com

A lot of publicly and commercially accessible databases contain information about chemical structure and biological activity of drug-like organic compounds [1]. Several methods have been suggested to reduce inconsistency in publicly available bioactivity databases [1, 2]. Typically, these approaches are based on selecting the compounds investigated by a single team of authors to reduce the impact of different assays on the activity measurements. However, there is still an issue how to create consistent data sets for the purposes of QSAR modeling using the large-scale databases of chemical compounds. In our study we investigated the ways to automatically prepare the modeling sets using the Integrity and ChEMBL databases as the examples of the commercially and publicly available databases respectively. We selected HIV-1 reverse transcriptase (RT) inhibitors for this research because this target provides a good case due to the presence of the multiple assays results in the databases. The structures of all HIV-1 RT inhibitors available from ChEMBL and Integrity were collected, including compounds assayed against both wild type and mutants of RT. Integrity provided a data set of 1,327 records (564 unique compounds) tested in more than 1300 bioassays approximately. ChEMBL yielded 3,787 records (2,297 unique compounds) tested in about 100 bioassays. For each of two general subclasses of HIV-1 (wild type of RT and the mutant forms of RT) we suggested several different ways to compile data sets for creating QSAR models: (1) selection of all compounds tested against a specific end-point; (2) selection of the compounds tested using one method and material (biological assay); (3) selection of the compounds derived from one scientific publication. We used a program GUSAR to build QSAR models. We tested the performance of the obtained QSAR models with leave 30% out cross-validation (LMO) and five-fold cross validation procedures; we then discussed the compatibility of the data from ChEMBL and Integrity.

For the most of modelling sets from Integrity database we observed an increase of the performance of the QSAR models created by the second compiling method in comparison to the first one (characteristics of the best model from the second compiling method, the data set “Antigen assay, Mononuclear cells (blood) as a material”: N=52; R^2=0.85; Q^2=0.76; F=7.7; SD = 0.91; R^2_{LMO}=0.75; R^2_{5fold} =0.64). However for the data sets from ChEMBL we did not observe similar trend. We have suggested this is a result either of the insufficient annotation or of the incomplete description of the assays in the scientific publication, which lead to the very fuzzy classification of the assay types in ChEMBL that does not make sense in the terms of the consistency. That observation corresponds to the conclusions of Kaliokoski et al. [1] We could not create the modelling sets using the third compiling method (selection of the compounds derived from one scientific publication) for the data sets from Integrity database, while the third compiling method leads to the increase of the performance of the models built on the data sets from ChEMBL database. We also proposed an algorithm to automatically match data from ChEMBL and Integrity on the compounds that were tested in the similar experimental conditions.

1. Kalliokoski T. et al. *PloS One*, 2013, **8**: e61007.
2. Muresan S. et al. *Drug Discov. Today*, 2011, **16**, 1019–1030.

This work was supported by the Russian Foundation of Basic Research (grant No. 13-04-91455_NIH-a).