HOW TO INCREASE THE CONCORDANCE OF THE EXPERIMENTAL DATA FOR QSAR MODELING: CASE STUDY FOR HIV-1 REVERSE TRANSCRIPTASE INHIBITORS

O.A.Tarasova[1], I. Mayzus[2], A. Rzhetsky[2], D.A. Filimonov[1], V.V. Poroikov[1]

[1] Institute of Biomedical Chemistry, Moscow, Russia;
[2] Computation Institute, Institute for Genomics and Systems Biology University of Chicago, Chicago, IL 60637, USA

Large-scale publicly and commercially accessible databases and scientific publications are important sources of the training sets for the analysis of quantitative structure-activity relationships (QSAR). The data contained in these sources may have the low concordance due to the significant differences in activity values for the same compound obtained in different studies [1]. To avoid the decrease in accuracy of the appropriate QSAR models, several problems should be analyzed. First, how to use the data from the databases to create accurate and predictive QSAR models. Second, whether the data from different databases might be mixed and matched. Third, whether we can use the data from the scientific publications to increase the concordance of the experimental data significantly and how text-mining algorithms can help to mix and match experimental data from these sources.

We have investigated the suitability of commercially and publicly available databases to QSAR modeling of antiviral activity (HIV-1 reverse transcriptase (RT) inhibition) [2]. We presented several methods for the creation of modeling (i.e., training and test) sets from two, either commercially or freely available, databases: Thomson Reuters Integrity and ChEMBL. We found that the performances of QSAR models obtained using these different modeling set compilation methods differed significantly from each other. Compound sets aggregated by target only typically yielded poorly predictive models. The best results were obtained using training sets compiled for compounds tested using only one particular type of assay performed using the specific biological material. One of the limitations of the “mix-and-matching” assay data across ChEMBL and Integrity is the general lack of complete and semantic/computer-parsable descriptions of assays methodology carried by the databases of these two investigated biologically active compounds that would allow one to determine mix-and-matchability of the resulted sets at the assay level.

In the present study, we propose the development and testing a text-mining algorithm dedicated to the creation of the highly homogeneous modeling sets, and illustrate why this procedure may be further used for the building of the QSAR models with high performance. In particular, the suggested approach includes: (1) automated selection of the relevant scientific publications, which contain the details of the biological assays; (2) automated selection of the fragments of the texts of the scientific publications with the subsequent description of the biological assays; (3) preparation of the datasets containing the feature vectors that are arranged according to the particular biological assay method; (4) evaluation of the similarity between the feature vectors obtained for different biological assays and (5) prediction of the belonging to the class associated with the subsequent biological assay method based on the random forest classifier and deep neural networks.

Our approach was approved for the creation of the highly concordant modeling sets of the HIV-1 RT inhibitors. Over 150 relevant scientific publications were retrieved from PubMed database in PDF format, transformed from PDF to plain text using PDFLib TET tool and processed as described earlier using the series of Python (2.6) scripts and the JAVA libraries available in Lingpipe 4.1.0. Preliminary testing of the method allowed obtaining the mean balanced accuracy of prediction 86%. The results of prediction and further QSAR models creation will be discussed.

1. S. Muresan et al., Drug Discovery Today., 2011, 16 (23-24), 1019−1030.
2. O. Tarasova et al. J. Chem. Inf. Model., 2015, 55 (7), 1388-1399.