The information contained in the large databases (DB) of chemical compounds, which represent an important source of data sets for QSAR modeling, is characterized by the variability of activity values measured in different experiments and/or by different laboratories [1]. Commercially available database that includes the data on the materials and methods of biological testing in the structured form is Thomson Reuters Integrity DB [2]. The largest freely available database of chemical compounds is CHEMBL database [3].

The purpose of our study is to investigate how the inconsistency of heterogenic data in the training sets influences on the quality of QSAR models that may be obtained, and propose the approach to improve the accuracy and predictivity of the models based on the data sets from INTEGRITY and CHEMBL databases.

METHODS

We used data sets of reverse transcriptase (RT) inhibitors extracted from the Integrity and CHEML DB to study the quality of QSAR models built for the data sets associated to different methods and materials of biological testing. When multiple IC50 values characterized the reverse transcriptase inhibiting activity were available in the database for the same compound, median values were used for creating the QSAR models. QSAR models were built using computer program GUSAR (General Unrestricted Structure-Activity Relationships) [4, 5], which superiority was shown in comparison with several other popular methods [4]. The accuracy and predictivity of the obtained QSAR models was estimated using leave 30% out cross-validation (LMO) and y-randomization (y-rand) procedures.

RESULTS

QSAR models built on the basis of the total heterogenous datasets included all compounds that were tested on RT inhibiting activity have very poor accuracy and predictivity (Figures 1, a, b, the first column). We developed a general automated workflow, which allows splitting data extracted from the Integrity database onto sub-sets grouped according to the material and methods of testing (INTEGRITY) and source of the data (INTEGRITY, CHEMBL). We propose that comparatively low quality of the models for the majority of combinations (material/method of testing) can be explained by small amount of the compounds collected from a single source of the data (i.e. small amount of the compounds of the single chemical class tested at the same experimental conditions). The specific conditions of the method of testing can influence on the accuracy of the QSAR models as well. The models with the best accuracy and predictivity were obtained for data sets corresponding to the radioactivity assay (PCR-based method), collected from the INTEGRITY database. The best results for the sub-sets corresponding to the cell-based assay were obtained for the data set corresponding to the antigen assay using mononuclear cells (blood), human (phosphohemagglutinin-stimulated).

Twenty three models were built using sub-sets from CHEMBL data sets obtained according to the certain source of biological information (particular paper). We revealed that such preprocessing of the data allows obtaining better models in comparison with usage of CHEMBL data without any division onto the subsets (the best model: F2 = 0.99,Q2 = 0.61, LMO = 0.64, y-rand = -0.24, (data set (N=56) corresponds to the data source Pubmed_ID:7683054 ).

The combination of the data from CHEMBL and INTEGRITY database was not analyzed due to the different data source used mainly and absence of the structured data about material and method of biological testing in CHEMBL DB. The models corresponded to the data sets from a single source of data were not created for the sets extracted from the INTEGRITY DB due to the lack of data.

CONCLUSIONS

An automatic procedure has been developed to obtain more homogeneous and consistent data sets for QSAR modeling based on the data sets from large scale databases. QSAR models obtained with more consistent and homogeneous training sets have better accuracy and predictivity. The quality of models is higher for some sets collected from INTEGRITY DB that corresponds to a specific pair material and method in comparison to the majority of the data sets collected from CHEMBL. Data sets from CHEMBL are characterized by the greater consistency of data in terms of chemical similarity since the set of compounds corresponding to single data source (usually over 10 structures) have high similarity in comparison to one-two typical chemical structure(s) from a single data source available in the INTEGRITY. The combining of data from CHEMCL and INTEGRITY is an open problem due to the absence of the unified terminology of the material and method of biological testing of chemical compounds. Therefore the scientific community needs to join forces and to perform an integration of the data from different sources using unified terminology.