DOI: 10.1002/minf.201000151

QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction

Alexey Lagunin,*[a] Alexey Zakharov,[a] Dmitry Filimonov,[a] and Vladimir Poroikov[a]

Presented at the 18th European Symposium on Quantitative Structure Activity Relationships, EuroQSAR 2010, Rhodes, Greece

Abstract: The method for QSAR modelling of rat acute toxicity based on the combination of QNA (Quantitative Neighbourhoods of Atoms) descriptors, PASS (Prediction of Activity Spectra for Substances) predictions and self-consistent regression (SCR) is presented. PASS predicted biological activity profiles are used as independent input variables for QSAR modelling with SCR. QSAR models were developed using LD_{50} values for compounds tested on rats with four types of administration (oral, intravenous, intraperitoneal, subcutaneous). The proposed method was evaluated on the set of compounds tested for acute rat toxicity with oral administration (7286 compounds) used for testing the known QSAR methods in T.E.S.T. 3.0 program (U.S. EPA). The several other sets of compounds tested for acute rat toxicity by different routes of administration selected from SYMYX MDL Toxicity Database were used too. The method was compared with the results of prediction of acute rodent toxicity for noncongeneric sets obtained by ACD/Labs Inc. The test sets were predicted with regards to the applicability domain. Comparison of accuracy for QSAR models obtained separately using QNA descriptors, PASS predictions, nearest neighbours' assessment with consensus models clearly demonstrated the benefits of consensus prediction. Free available web-service for prediction of LD_{50} values of rat acute toxicity was developed: http://www.pharmaexpert.ru/GUSAR/AcuToxPredict/

Keywords: QSAR · Acute rodent toxicity · QNA · PASS · SCR · Prediction · Pathways

1 Introduction

Estimation of rodent acute toxicity is an important task in drug design and risk assessment of chemicals. Experimental testing of compounds on rodent acute toxicity being costly is also criticized on ethical reasons. The European Community Regulation on chemicals and their safety use (REACH), which has been started at 2007, anticipates the development of computer-aided methods for the analysis of "structure-activity" relationships (e.g. IUCLID) and the study of toxic effects for several dozen thousands of chemical substances. Following this trend, we have developed a new method for rodent acute toxicity QSAR modelling realized in GUSAR software [1]. This method is based on the combination of QNA (Quantitative Neighbourhoods of Atoms) descriptors [1], PASS (Prediction of Activity Spectra for Substances) predicted biological activity profiles [2–4], and self-consistent regression [1,5].

Acute toxicity is considered as the adverse effects occurring within a given time, following a single exposure to a substance [6]. LD_{50} value is one of important characteristics of acute toxicity that corresponds to the dose causing 50% mortality within 24 hours of administration. Acute oral, dermal and inhalation rodent toxicity are important parameters for general toxicological risk assessments, whereas oral, intraperitoneal and intravenous acute rodent toxicity are important in drug design. Mice and rats are the main species used in these studies. There is a lot of LD_{50} data for mice and rats available in literature and databases [7–9]. The necessity of in silico estimation of LD_{50} values led to creation and application of different SAR and QSAR methods. These methods were recently reviewed [10,11].

Acute toxicity is a complex phenomenon which includes action of chemicals through different biochemical mechanisms. Such complexity hampers the process of QSAR modelling and leads to moderate accuracy of prediction for noncongeneric sets. Nevertheless, several works with reasonable results appeared in the past two years [10–12]. The advantages of consensus prediction have been demonstrated before [13,14]. It is considered that the consensus models decrease the variance of individual models. All individual models contain varying amounts of predictions with uncertainty and their averaging leads to more reliable results [15]. In these works a combination of different QSAR methods were used for the achievement of reasonable re-

[a] *A. Lagunin, A. Zakharov, D. Filimonov, V. Poroikov*
Department for Bioinformatics, Institute of Biomedical Chemistry
of the Russian Academy of Medical Sciences
Pogodinskaya Str., 10, Moscow, 119121, Russia
phone/fax: +7 499 2553029/+7499 2450857
**e-mail: alexey.lagunin@ibmc.msk.ru*

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201000151.

sults. In this study the accuracy and predictability of novel QSAR approach for consensus prediction of rat acute toxicity in comparison with other methods were analyzed. Utilization of PASS predicted biological activity profiles as the basis for QSAR modelling, provides the possibility for biological interpretation of the models, that corresponds to the OECD recommendations for QSAR models.

2 Materials and Methods

2.1 QSAR Modelling on the Basis of QNA Descriptors

QSAR modelling on the basis of QNA descriptors has been previously implemented in the GUSAR software [1]. Reasonable results obtained by GUSAR modelling for different biological endpoints [1] suggest the possibility of using this method to the modelling of acute rodent toxicity. More detailed explanation of the approach is represented in Supporting Information. It is briefly described below.

QNA descriptors are calculated based on the connectivity matrix (C), standard values of ionization potential (IP) and electron affinity (EA) of atoms in a molecule [1]. QNA describes each atom in a molecule, and, at the same time, each *P* and *Q* values depend on the whole composition and structure of a molecule. The estimation of target property of chemical compound is calculated as the mean value of the function of *P* and *Q* in the points of the atoms of a molecule in QNA descriptors' space. We have proposed to use two-dimensional Chebyshev polynomials for approximation of this function of *P* and *Q*, so, the independent regression variables are calculated as average values of particular two-dimensional Chebyshev polynomials of *P* and *Q* for molecule atoms.

QNA descriptors and their polynomial transformations do not provide information on the shape and volume of a molecule although this information may be important for determination of the structure-activity relationships. Therefore, these parameters were added to the obtained from Chebyshev polynomials variables. Topological length of a molecule is the maximal distance calculated in the number of bonds between any two atoms (including hydrogen), and the volume of a molecule as the sum of each atom's volume.

The number of initial variables for rodent acute toxicity QSAR modelling depends on the number of compounds in the training set and corresponds to the number of Chebyshev polynomials plus the number of the first, second and third power of topological length and the volume of a molecule. If the number of compounds in the training set varied from 100 to 2000, then the number of initial variables was one-half of the number of compounds in the training set. If the number of compounds in the training set exceeds 2000, then the number of initial variables is 1000.

GUSAR algorithm generates three types of QSAR models based on QNA descriptors: QNA descriptors are calculated for all atoms or for only those atoms in a molecule, which have two or more immediate neighbours; the coefficient before the connectivity matrix and the parameters of Chebyshev polynomials are changed. The detailed algorithm is described in the Supporting Information, Part 1. The final QSAR model is the consensus of built in this way different QNA based models.

2.2 QSAR Modelling on the Basis of PASS Predicted Biological Activity Profiles

The current version of PASS (10.1) predicts 4130 types of biological activity with the mean prediction accuracy of about 95%. Currently, the list of predictable biological activities includes 501 pharmacotherapeutic effects, (e.g., Antihypertensive, Hepatoprotectant, Nootropic, etc.), 3295 mechanisms of action, (e.g., 5 Hydroxytryptamine antagonist, Acetylcholine M1 receptor agonist, Cyclooxygenase inhibitor, etc.), 57 adverse & toxic effects (e.g., carcinogenic, Mutagenic, Hematotoxic, etc.), 199 metabolic terms (e.g., CYP1A inducer, CYP1A1 inhibitor, CYP3A4 substrate, etc.) 49 transporter proteins (e.g., P-glycoprotein 3 inhibitor, Nucleoside transporters inhibitors) and 29 activities related to gene expression (e.g., TH expression enhancer, TNF expression inhibitor, VEGF expression inhibitor). The detailed description of PASS algorithm, including the list of predictable activities, is represented in [2–4] and in Supporting Information (Part 2 and 4). The results of PASS prediction are given as a list of biological activities, for which the difference between probabilities to be active (*Pa*) and to be inactive (*Pi*) was calculated.

The *Pa-Pi* values for activities randomly selected from the total list of predicted biological activities were used as input independent variables for regression analysis, to obtain different QSAR models. Similar to the QSAR analysis with QNA descriptors (Section 2.1), topological length and volume of molecules were added as variables to the profiles of biological activity; the number of initial variables for creating regression models was also selected depending on the number of compounds in the training set.

2.3 Self-Consistent Regression

GUSAR uses self-consistent regression for models building. Self-consistent regression (SCR) is based on the regularized least-squares method [1,5].

Unlike the stepwise regression and other methods of combinatorial search, the initial SCR model includes all regressors. The basic result of SCR method is the removal of variables, which are worse for the description of appropriate value [1,5]. The number of the final variables in QSAR equation selected after the self-consistent regression procedure is significantly less compared to the number of the initial variables. Nevertheless, the final model contains the set of variables, correctly representing the existing relationship.

2.4 Nearest Neighbour's Correction

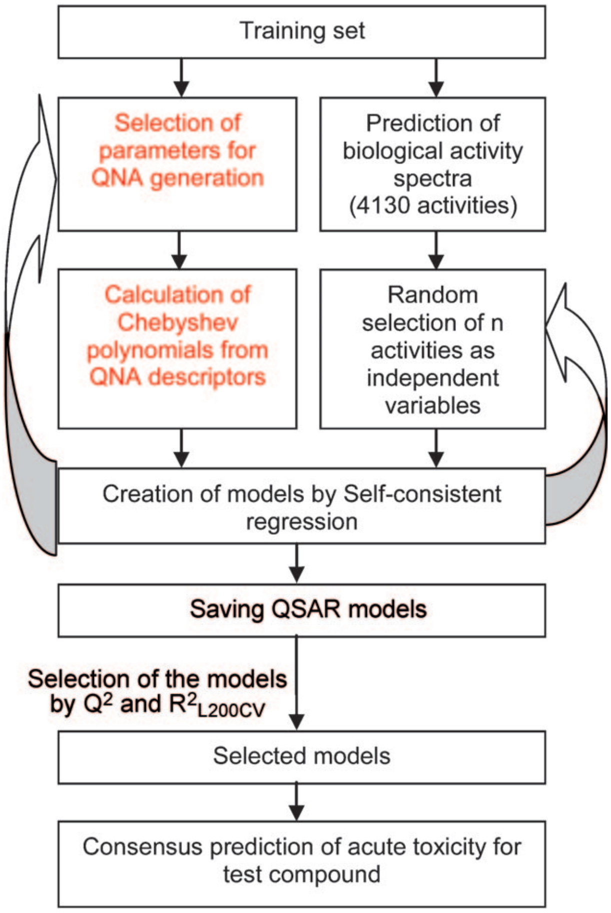
It is well known that the joint use of global and local models for noncongeneric sets improves the quality of QSAR models [16]. We used the experimental data of three nearest neighbours (NN) for correction of prediction value obtained from the regression model. Similarity of any chemical compounds' pairs is estimated as Pearson's coefficient calculated in the space of independent variables obtained after SCR. The mean experimental LD_{50} value obtained for three nearest neighbour compounds from the training set was averaged with the predicted LD_{50} value of the test compound.

2.5 Applicability Domain

Pearson's coefficient as a measure of pairwise similarity of chemical compounds was calculated in the space of independent variables obtained after SCR. Using these estimates, three nearest neighbouring compounds in the training set similar to the analyzed structure were found. The average similarity with these three compounds was used for the assessment of applicability domain (AD) of the model. If the average similarity exceeds the threshold then the chemical compound under prediction falls in AD of the model and vice-versa. The higher value of the threshold was selected the more similar compounds fell in AD of the model. In this study we investigated several thresholds for AD: 0.7, 0.8 and 0.9.

2.6 Consensus Modelling

The final predicted value is estimated including a weighted average of predicted values from each obtained QSAR model (QSAR models provide the predictions that are within the respective applicability domains). Predicted value obtained from each developed model is weighted on similarity value calculated in estimating the applicability domain. The algorithm combining the results of QSAR modelling on the basis of QNA descriptors and PASS predicted biological activity profiles, is represented in Scheme 1.


Training set
Selection of parameters for QNA generation
Prediction of biological activity spectra (4130 activities)
Calculation of Chebyshev polynomials from QNA descriptors
Random selection of n activities as independent variables
Creation of models by Self-consistent regression
Saving QSAR models
Selection of the models by Q^2 and R^2_{L200CV}
Selected models
Consensus prediction of acute toxicity for test compound

Scheme 1. Creation of consensus model.

2.7 Data on Acute Rat Toxicity

2.7.1 Data from SYMYX MDL Toxicity Database

We used in-house database created on the basis of data from SYMYX MDL Toxicity Database [7]. It included the information about ~10000 chemical structures with the data on acute rat toxicity. The toxicity end-points are based on the log10 representation of LD_{50} values (mmol/kg) for the rats with four types of administration: oral, intraperitoneal, intravenous and subcutaneous. In the preparation of studied sets, all data were reviewed to remove any salts, mixtures, inorganic compounds and polymers. The quality of data is important for creation of accurate QSAR models. We made special selection of the data before creation of QSAR models. Some compounds had more than one LD_{50} value for an appropriate type of administration. For such compounds we used an average value, but if the deviation of LD_{50} values was higher 0.5 log10 then such compounds are excluded from the further analysis. The deviations between the data for different types of administration were also reviewed. All compounds with deviations of LD_{50} values between any pairs of routes of administration exceeded 2× *RMSD* (Root Mean Square Deviation) were excluded. The compounds with LD_{50} values that were out the interval [mean LD_{50} value $\pm 5 \times SD$] were also excluded for each type of administration (*SD* – Standard Deviation). Thus, about 400 compounds were excluded. Table 1 displays the numbers of compounds, mean values, intervals and *SD* for the studied sets.

2.7.2 Data from EPA Dataset

We have compared our approach with well-known methods realized in T.E.S.T. 3.0 program (Toxicity Estimation Software Tool) provided by U.S. EPA (Environmental Protection

Table 1. Characteristics of the sets selected on the basis of SYMYX MDL Toxicity Database.

Administration	*N*[a]	Mean value[b]	Intervals[c]	*SD*[d]
Oral	8972	0.481	[−3.9 : 2.7]	0.913
Intraperitoneal	3549	0.023	[−4.1 : 3.5]	0.893
Intravenous	1314	−0.524	[−4.6 : 2.5]	0.981
Subcutaneous	1084	0.177	[−4.9 : 2.3]	1.042

[a] Number of compounds; [b] Mean value of Log10(LD_{50}) values in the set (mmol/kg); [c] Intervals between the minimal and maximal Log10(LD_{50}) values in the set, mmol/kg; [d] Standard deviation of Log10(LD_{50}) values in the set (mmol/kg).

Agency) [17]. This program includes models obtained using several QSAR methods:

1. *Hierarchical* method: The toxicity for a given query compound is estimated using the weighted average of the predictions from several different models. The different models are obtained by Ward's method to divide the training set into a series of structurally similar clusters. A genetic algorithm based technique is used to generate models for each cluster.

2. *FDA* method: The prediction for each test chemical is made using a new model that fits to the chemicals, which are most similar to the test compound.

3. *Nearest neighbour* method: The predicted toxicity is estimated by calculating an average value for three chemicals in the training set that are most similar to the test chemical.

4. *Consensus* method: The predicted toxicity is estimated by the average of the predicted toxicities from the above QSAR methods (provided the predictions are within the respective applicability domains).

Oral rat LD_{50} data set was used for comparing the proposed method with those realized in T.E.S.T. 3.0 program. This data set contained 7286 chemicals, which were randomly divided to the training and test set by the authors of T.E.S.T. 3.0 program. The modelled endpoint was −Log10(LD_{50} mol/kg). The training set includes 5828 chemical compounds. The test set includes 1458 chemical compounds. The same training and test sets were used in this comparative study.

3 Results and Discussion

To evaluate the accuracy of prediction of the proposed method and reveal the impact of each algorithm, we performed the following experiment using four sets with LD_{50} values studied on rats (oral, intravenous, intraperitoneal and subcutaneous routes of administrations) in mmol/kg. The sets were randomly divided onto the training and test sets in proportion of 70 % and 30 %, respectively, to compare our models with the results given by Sazonovas et al. [11]. Leave-twenty%-out cross-validation procedure was carried out twenty times dividing of the initial training sets on the training and test sets into the proportion of 80 % and 20 %, respectively ($R^2_{L20\%Out}$). Forty QSAR models on the basis of PASS predictions and forty QSAR models on the basis of QNA descriptors were created with the training set prepared in accordance with the route of administration. Three types of consensus prediction were obtained for the appropriate test sets:

1. Consensus from all models (both QNA and PASS prediction results) created for the appropriate route of administration (PASS&QNA);

2. Consensus from the models on the basis of PASS prediction results created for the appropriate route of administration (PASS);

3. Consensus from the models on the basis of QNA descriptors created for the appropriate route of administration (QNA).

The mean values of characteristics of the created QSAR models and their validation are represented in Table 2. The plots with the observed versus predicted values for the test sets according to the route of administration are given in Figures 1–4.

From these data one may conclude that QSAR models have a reasonable quality. Some points in the Figures 1–4 can be considered as outliers, but their number is negligi-

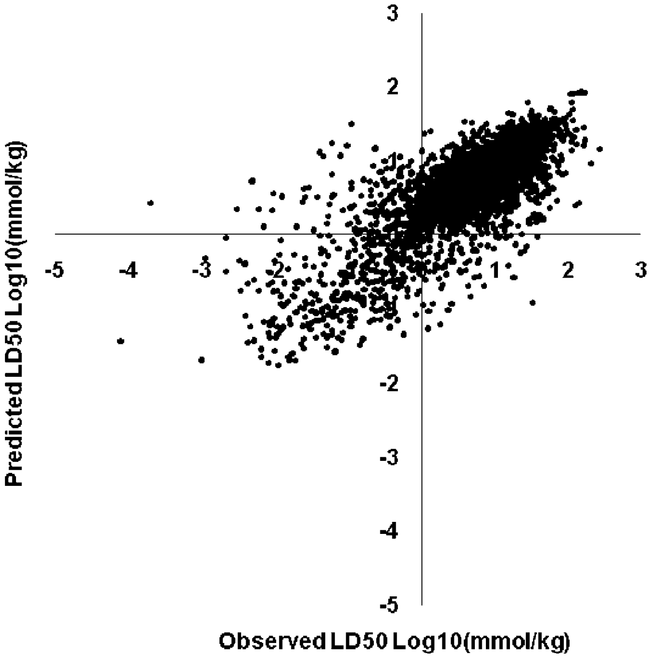


Figure 1. Observed versus predicted LD_{50} values for test set at oral administration.

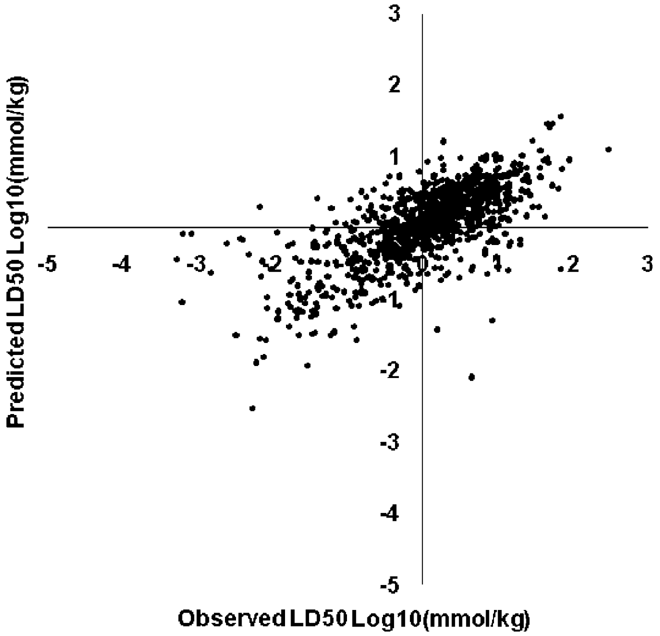


Figure 2. Observed versus predicted LD_{50} values for test set at intraperitoneal administration.

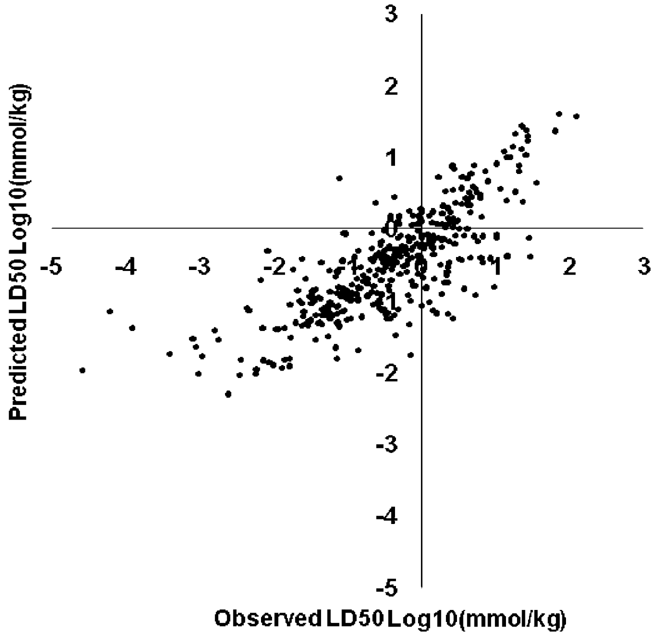


Figure 3. Observed versus predicted LD_{50} values for test set at intravenous administration.

ble compared with the total number of all points (~4,500). In general, Figures 1–4 show that there are no any visible artefacts in the predicted values.

The table shows that in most cases QSAR models created on the basis of PASS predicted biological activity profiles reveal the highest accuracy of prediction in comparison with the models created on the basis of QNA descriptors. It is also clear that the use of consensus between QSAR models created on PASS prediction results and QNA descriptors leads to the increase of both the accuracy of prediction (R_{test}^2 and/or $RMSE_{test}$) and the coverage of test sets (in all cases). Coverage means the percent of compounds from the test set in Applicability Domain. The mean increase of the coverage was 10.3%. Correction by the data of the nearest neighbours is also very important for achievement of high accuracy. The use of the nearest neighbours' correction increased R_{test}^2 value in average on 0.06 and decreased $RMSE_{test}$ value in average on 0.035. Though the speed of prediction for consensus models decreases, it remains sufficiently high for the practical use. Even in the worst case (prediction of LD_{50} value for oral type of administration) the speed of prediction was 0.78 compounds per second, which means that prediction for about 2850 compounds is calculated within an hour. The speed of prediction was obtained on PC with Core-i7 920 CPU, 6 GB RAM DDR3 1033 and Windows 7.

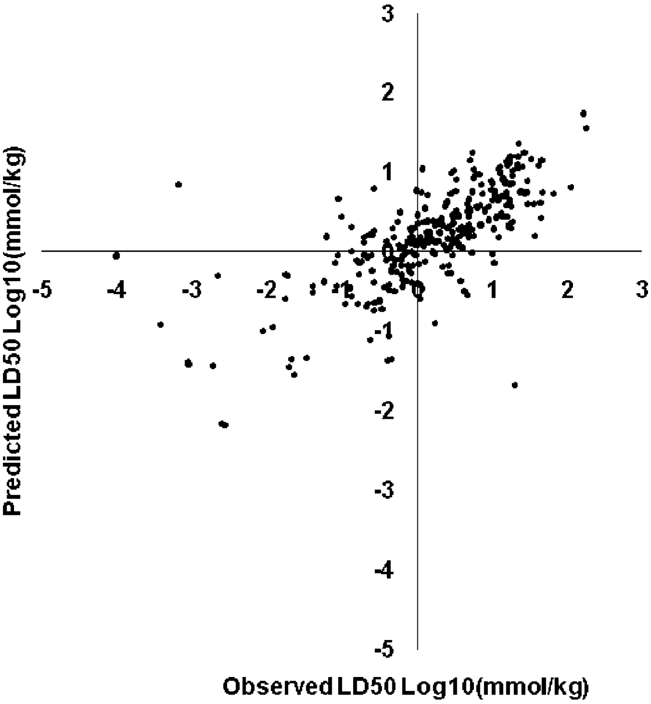


Figure 4. Observed versus predicted LD_{50} values for test set at subcutaneous administration.

Table 2 shows that the ratio of the mean number of variables to the number of compounds in the training sets is less than 1:5. Q^2 and R_{test}^2 have comparable values in all cases; their difference between R^2 and Q^2 values is less than 0.1. It means that 1) the created models are not overfitted; 2) the created QSAR models are robust.

It was shown that leave-20%-out (L20%Out) cross-validation procedure can successfully be applied for assessment of predictivity and robustness of models [1]. For example, in case of intravenous type of administration eighty QSAR models were generated. Leave-20%-out cross-validation procedure was performed 20 times for each model. The average value of $R^2_{L20\%Out}$ obtained for eighty QSAR models is 0.428 and standard deviation is 0.026 (Supporting Information, Part 3). These results show that despite random selection of initial variables, the developed QSAR method is robust and reproducible, because all obtained models have a comparable predictivity. The same results (not shown) were obtained for the sets with the other type of administration.

We also compared the results of the proposed approach with those of Sazonovas et al. [11] based on the combination of general and local QSAR models and calculated reliability index for prediction of rodent acute toxicity. It is obvious that direct comparison is not possible because of different training and test sets used [11]. Nevertheless, the overall depiction can obtained due to the similar of data sources (SYMYX MDL Toxicity Database includes RTECS data) and comparable sizes of the training and test sets. Thus, one can expect similar chemical space and activity distribution. Table 2 shows close accuracy of prediction obtained by both methods. In all cases the proposed approach has higher value of the coverage. The results of consensus modelling demonstrate that the values of $R_{test}{}^2$ are equal or exceed 0.5 and $RMSE_{test}$ are less 0.7.

We also studied how the increase of threshold for applicability domain influences the accuracy of prediction and coverage of the sets. For this purpose, the test sets for each type of administration were predicted by the appropriate best consensus models with different values of AD (0.7, 0.8, and 0.9). Table 3 shows the results of this study.

Table 2. Characteristics of QSAR models for the sets from SYMYX MDL Toxicity Database (the number in brackets is the number of models based on the appropriate descriptors).

Models	$R^{2[c]}$	$Q^{2[d]}$	$F^{[e]}$	$SD^{[f]}$	$V^{[g]}$	$R^2_{L20\%O}{}^{[h]}$	$R_{test}{}^2$	$RMSE_{test}$	Coverage (%)[i]	Compds/s[j]
Oral	SYMYX MDL Toxicity Database ($N_{training}{}^{[a]} = 6280$; $N_{test}{}^{[b]} = 2692$)									
PASS(35)	0.61	0.57	16.654	0.576	420	0.44	0.53	0.59	92.4	0.85
PASS(35)&NN	0.61	0.57	16.654	0.576	420	0.44	0.59	0.56	92.4	0.85
QNA(5)	0.54	0.50	16.852	0.622	328	0.36	0.48	0.62	88.8	6.97
QNA(5)&NN	0.54	0.50	16.852	0.622	328	0.36	0.54	0.58	88.8	6.97
PASS(35)&QNA(5)	0.61	0.57	16.685	0.573	408	0.40	0.53	0.60	97.5	0.79
PASS(35)&QNA(5)&NN	0.61	0.57	16.685	0.573	408	0.40	0.59	0.57	97.5	0.79
Best results of Sazonovas et al. ($N_{training}{}^{[a]} = 6464$; $N_{test}{}^{[b]} = 2167$)										
ACD/Labs	n/a[k]	n/a	n/a	n/a	n/a	n/a	0.56	0.59	91.0	n/a
Intraperitoneal	($N_{training}{}^{[a]} = 2480$; $N_{test}{}^{[b]} = 1065$)									
PASS(40)	0.61	0.52	6.144	0.534	332	0.325	0.39	0.60	92.7	2
PASS(40)&NN	0.61	0.52	6.144	0.534	332	0.325	0.48	0.55	93.0	2
QNA(28)	0.60	0.50	4.965	0.544	320	0.266	0.36	0.61	92.0	2.6
QNA(28)&NN	0.60	0.50	4.965	0.544	320	0.266	0.400	0.60	92.0	2.6
PASS(40)&QNA(28)	0.66	0.56	6.029	0.518	327	0.301	0.42	0.60	98.3	2.2
PASS(40)&QNA(28)&NN	0.66	0.56	6.029	0.518	327	0.301	0.48	0.57	98.3	2.2
Best results of Sazonovas et al. ($N_{training}{}^{[a]} = 3751$; $N_{test}{}^{[b]} = 1251$)										
ACD/Labs	n/a	n/a	n/a	n/a	n/a	n/a	0.42	0.58	90.0	n/a
Intravenous	($N_{training}{}^{[a]} = 920$; $N_{test}{}^{[b]} = 394$)									
PASS(40)	0.72	0.64	9.848	0.529	142	0.45	0.57	0.65	95.4	7.4
PASS(40)&kNN	0.72	0.64	9.848	0.529	142	0.45	0.60	0.63	95.4	7.4
QNA(10)	0.66	0.58	9.457	0.577	123	0.42	0.55	0.65	93.4	28.1
QNA(10)&kNN	0.66	0.58	9.457	0.577	123	0.42	0.59	0.62	93.4	28.1
PASS(40)&QNA(10)	0.73	0.66	9.964	0.524	138	0.44	0.59	0.65	99.2	5.9
PASS(40)&QNA(10)&kNN	0.73	0.66	9.964	0.524	138	0.44	0.63	0.62	99.2	5.9
Subcutaneous	($N_{training}{}^{[a]} = 759$; $N_{test}{}^{[b]} = 325$)									
PASS(5)	0.66	0.55	5.432	0.612	130	0.32	0.40	0.73	88.6	54.2
PASS(5)&NN	0.66	0.55	5.432	0.612	130	0.32	0.49	0.67	88.9	54.2
QNA(2)	0.60	0.50	5.432	0.612	130	0.27	0.42	0.72	47.7	325.0
QNA(2)&NN	0.60	0.50	5.483	0.645	122	0.27	0.50	0.67	48.3	325.0
PASS(5)&QNA(2)	0.69	0.59	5.484	0.596	128	0.30	0.41	0.75	91.7	40.6
PASS(5)&QNA(2)&NN	0.69	0.59	5.484	0.596	128	0.30	0.50	0.69	92.0	40.6

[a] Number of compounds in the training set; [b] Number of compounds in the test set; [c] Average R^2 of the models calculated for the appropriate training set; [d] Average Q^2 of the models calculated for the appropriate training set; [e] Fisher coefficient; [f] Standard deviation; [g] Number of independent variables in the model; [h] Results of 20% out cross-validation procedure calculated for the appropriate training set; [i] % compounds from the test set in Applicability Domain; [j] Number of compounds for that prediction was assisted per second (speed of prediction). It was calculated on PC with Core-i7 920 CPU, 6Gb RAM DDR3 1033 and Windows 7; [k] not available.

Table 3. Accuracy of prediction for the test sets at different threshold of applicability domain.

	R_{test}^{2}[a]	$RMSE_{test}$[b]	Coverage (%)[c]
Oral			
AD > 0.7	0.585	0.567	97.4
AD > 0.8	0.587	0.567	95.6
AD > 0.9	0.611	0.550	84.7
Intraperitoneal			
AD > 0.7	0.479	0.573	98.3
AD > 0.8	0.585	0.567	96.1
AD > 0.9	0.510	0.543	78.6
Intravenous			
AD > 0.7	0.625	0.620	99.2
AD > 0.8	0.622	0.620	99.0
AD > 0.9	0.644	0.587	92.6
Subcutaneous			
AD > 0.7	0.500	0.689	92.0
AD > 0.8	0.503	0.711	81.2
AD > 0.9	0.628	0.598	52.9

[a] Average R^2 of the models calculated for the appropriate test set; [b] Average *RMSE* (Root Mean Square Error) of the models calculated for the appropriate test set; [c] % compounds from the test set in Applicability Domain.

The values of R_{test}^{2} depend on the data spread of those compounds which fail in AD and may change with the change of the coverage. Thus, $RMSE_{test}$ values represent more reliable estimation of accuracy prediction than the values of R_{test}^{2}. Table 3 shows that the increase of AD threshold leads to the decrease of $RMSE_{test}$ values and coverage. In this case more accurate prediction may be obtained by the reasonable decrease of coverage.

The proposed approach was also compared with the T.E.S.T. program (Toxicity Estimation Software Tool) Version 3.0, developed by U.S. Environmental Protection Agency, 2008 evaluated on the available training and test sets with the data on oral acute toxicity measured in LD_{50} (mmol/kg) values. The modelling results are represented in Table 4.

Table 4 demonstrates that the proposed method has the highest accuracy in comparison with the above mentioned methods and provides the highest speed of the prediction (18 times faster). As the training set contains compounds belonging to different chemical classes, we assume that the proposed algorithm provides accurate prediction even for heterogenic set of compounds.

Combination of the prediction results of both the proposed method and EPA consensus increased the accuracy of prediction and coverage of the test set (the last line in the table). R_{test}^{2} has increased from 0.639 to 0.670. $RMSE_{test}$ has decreased from 0.581 to 0.558. The coverage has achieved 100%.

One of the features of the proposed method is creation of equation on the basis of PASS prediction results. The use of PASS predicted biological activity profile in the search of the possible mechanisms of toxicity was proposed by Poroikov et al. in 2007 [18]. The application of PASS-predicted biological activity for creation of QSAR models of rat acute toxicity provides the revelation of probable mechanisms of action that might be the putative cause of acute toxicity. The example of QSAR equation for the best model of LD_{50} values for acute rat toxicity at subcutaneous route of administration is represented in Scheme 2.

The activities from the equations of the models for each type of administration have been analysed. The top 30 activities with the highest absolute frequency of occurrence in equations (40 equations were for each type of administration) are shown in Table 5.

The list of activity includes both the general biological activity directly related with toxicity of compounds (e.g. nephrotoxic; carcinogenic, mouse, female) and the actions on molecular targets that may also be related with toxicity. For example, inhibition of glucose-6-phosphate isomerase leads to hemolytic anemia and hematotoxic effect. Several activities related with regulation of inflammatory process (Leukotriene C4 antagonist; Leukotriene E4 antagonist; Prostaglandin EP2 agonist; Prostaglandin EP2 agonist; 12 Lipoxygenase inhibitor and Transforming growth factor antagonist). Stimulation of complement C5a convertase relates with blood coagulability. Inhibition of acetyl-CoA transferase influences the lipid regulation. Stimulation of α_{1a}-adrenoreceptor may cause hepatotoxicity [19]. Stimulation of adenosine A2b receptor causes vasodilatation. Interaction with 5 Hydroxytryptamine receptors may cause system neurological and cardiovascular effects. The most of pharmacological activities and chronic toxicities from the equation are not directly related to acute toxicity. It may be associative relationship between the prediction values of these activities and the acute toxicity values.

The selected activities may be used together with the data on the known biological pathways for the analysis of possible mechanisms of acute toxicity. An example of such analysis is represented in Figure 5. The list of activities related with the action on molecular targets was superposed on the known pathways by X-Platform software (Gene-

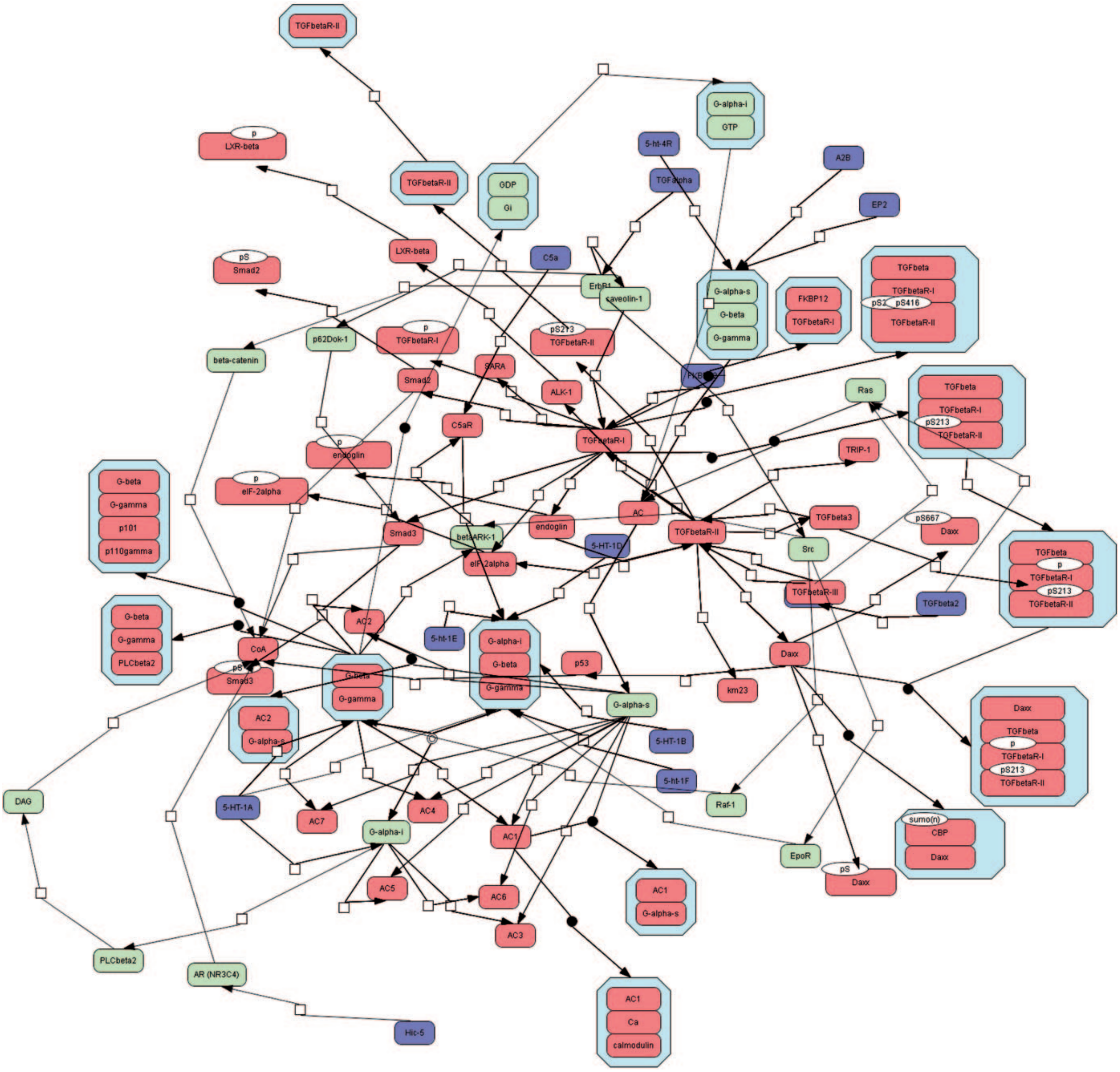


Figure 5. Relationships between the molecular targets related with activities in Table 5 (Selected targets are painted by blue).

Table 4. Comparison of the proposed approach with the results of T.E.S.T. 3.0 program (Toxicity Estimation Software Tool) (U.S. EPA, 2008).

Method	R_{test}^2	$RMSE_{test}$	Coverage (%)[b]	Compounds/s[c]
Hierarchical[a]	0.573	0.654	84.7	0.17
Nearest neighbour[a]	0.546	0.662	99.5	0.08
FDA[a]	0.555	0.658	98.7	0.33
Consensus (Hierarchical, Nearest neighbour, FDA)[a]	0.620	0.596	100	0.05
Consensus (PASS&QNA&NN)	0.639	0.581	95.2	0.90
Consensus (Hierarchical, Nearest neighbour, FDA)[a] + PASS&QNA&NN	0.670	0.558	100	0.04

[a] Given from User's Guide for T.E.S.T. (Toxicity Estimation Software Tool), a program to estimate toxicity from molecular structure, Version 3.0, U.S. Environmental Protection Agency, 2008; [b] % compounds from the test set in Applicability Domain; [c] Number of compounds predicted per second (velocity of prediction) calculated on PC with Core-i7 920 CPU, 6Gb RAM DDR3 1033 and Windows 7.

***LD*$_{50}$ Log10(mmol/kg)** = 0.212 × (Liver X receptor β-agonist) + 0.262 × (Hepatotoxic) + 0.352 × (Fibroblast growth factor 2 antagonist) − 0.177 × (CYP2C8 inducer) − 0.325 × (Actin polymerization inhibitor) − 1.21 × (Amidase inhibitor) + 0.411 × (Fucosterol-epoxide lyase inhibitor) + 0.957 × (Prokineticin receptor 1 antagonist) + 0.386 × (Muscular dystrophy treatment) − 3.98 × (Volume^2) + 0.0171 × (CYP2D6 inhibitor) − 0.228 × (Formate dehydrogenase inhibitor) + 0.859 × (Acetylcholine M3 receptor agonist) + 0.482 × (GP IIb/IIIa receptor antagonist) − 0.89 × (Arylformamidase inhibitor) − 1.1 × (Selectin L antagonist) − 0.741 × (Complement C5a convertase stimulant) + 0.864 × (Choleretic) − 1.12 × (Imidazoline I2 receptor antagonist) − 0.516 × (Proto-oncogene tyrosine-protein kinase Kit inhibitor) − 0.743 × (Cyclamate sulfohydrolase inhibitor) + 0.624 × (Catechol 1,2-dioxygenase inhibitor) + 0.26 × (ErbB-2 antagonist) + 1.25 × (Sphingosine 1-phosphate receptor 4 agonist) + 1.03 × (D-Lactate-2-sulfatase inhibitor) − 0.658 × (Glycerol-3-phosphate dehydrogenase inhibitor) − 2.62 × (Ribose-5-phosphate isomerase inhibitor) − 0.345 × (Sterol 24-C-methyltransferase inhibitor) + 0.593 × (Transforming growth factor β-antagonist) − 0.655 × (Interferon α-antagonist) + 0.945 × (Lactosylceramide α-2,3-sialyltransferase inhibitor) + 0.22 × (DNA directed RNA polymerase inhibitor) − 0.329 × (HIV-1 integrase × (3′-Processing) inhibitor) + 0.137 × (Anti-Helicobacter pylori) + 0.296 × (Antimycoplasmal) + 0.195 × (Carcinogenic, rat, female) − 1.56 × (UDP-*N*-acetylglucosamine-lysosomal-enzyme *N*-acetylglucosaminephosphotransferase inhibitor) + 1.36 × (Pyruvate, phosphate dikinase inhibitor) + 0.871 × (Queuine tRNA-ribosyltransferase inhibitor) + 0.24 × (Thrombopoietin agonist) − 0.127 × (Neuroprotector) − 0.433 × (Tyrosine kinase inhibitor) − 0.223 × (β_2-Adrenoreceptor agonist) − 1.09 × (CYP4A1 substrate) − 0.69 × (CYP4F8 substrate) + 0.205 × (Antipsoriatic) + 0.225 × (Interleukin 6 antagonist) + 1.09 × (Deoxycytidine deaminase inhibitor) + 4.01 × (Volume^2) − 2.86 × (Volume^2) + 0.294 × (Prostaglandin D2 agonist) − 0.339 × (Serine *O*-acetyltransferase inhibitor) + 0.519 × (Benzodiazepine receptor peripheral-type antagonist) + 0.795 × (Cortisone α-reductase inhibitor) + 0.15 × (Growth hormone releasing factor agonist) − 0.349 × (Antileukemic) + 0.271 × (Uric acid excretion stimulant) + 0.451 × (Collagenase 3 inhibitor) − 0.624 × (Lathosterol oxidase inhibitor) + 0.531 × (Mandelate 4-monooxygenase inhibitor) + 0.388 × (Acetate-CoA ligase inhibitor) − 0.233 × (5 Hydroxytryptamine 1 antagonist) − 0.303 × (Kidney function stimulant) − 0.172 × (Fibrinolytic) − 1.23 × (Ketosteroid monooxygenase inhibitor) + 0.673 × (Aldose 1-epimerase inhibitor) + 0.267 × (Homocitrate synthase inhibitor) + 0.271 × (Nitric oxide synthase inhibitor) + 0.274 × (Glycerol-1,2-cyclic-phosphate 2-phosphodiesterase inhibitor) − 0.373 × (Nitrate reductase × (cytochrome) inhibitor) − 0.5 × (Neuronal nitric oxide synthase inhibitor) + 0.139 × (Antiviral × (Hepatitis B)) + 0.37 × (Aspartate ammonia-lyase inhibitor) − 0.112 × (Antineoplastic × (head/neck cancer)) + 0.16 × (Cancer associated disorders treatment) − 0.416 × (CYP19 inhibitor) + 0.421 × (Folate antagonist) + 0.268 × (CYP2B5 substrate) + 0.156 × (Endocrine disorders treatment) − 0.283 × (Potassium channel × (ATP-sensitive) activator) − 0.171 × (Calpain inhibitor) − 0.194 × (Multiple inositol-polyphosphate phosphatase inhibitor) − 0.298 × (AMPA receptor agonist) + 0.363 × (Methylaspartate ammonia-lyase inhibitor) + 0.539 × (Plus-end-directed kinesin ATPase inhibitor) + 0.203 × (Signal transduction pathways inhibitor) + 0.265 × (*O*-aminophenol oxidase inhibitor) + 0.107 × (Gynecological disorders treatment) + 0.11 × (CYP1A inhibitor) + 0.0907 × (Age-related macular degeneration treatment) − 0.123 × (Postmenopausal disorders treatment) + 0.139 × (Antibiotic Glycopeptide-like) + 0.0764 × (Antiviral × (Hepatitis)) − 0.0792 × (Antioxidant) + 0.276 × (Hydroxymethylglutaryl-CoA lyase inhibitor) − 0.207 × (CYP1A1 inducer) + 0.0895 × (Sensitization) + 0.0834 × (Antibiotic Oxazolidinone-like) + 0.107 × (CYP1A2 inducer) + 0.0731 × (Antibiotic Trimethoprim-like) − 0.175 × (MAO-B substrate) + 0.145 × (Thyroid hormone β-agonist) − 0.0937 × (Thyroid hormone β_1-agonist) + 0.0975 × (Motilin receptor antagonist) + 0.149 × (Calcitonin gene-related peptide 1 receptor antagonist) + 0.111 × (Peptide α-*N*-acetyltransferase inhibitor) − 0.0454 × (Epoxide hydrolase 2 inhibitor) − 0.0727 × (Oxytocin agonist) − 0.0392 × (Carboxycyclohexadienyl dehydratase inhibitor) − 0.088 × (17-β-hydroxysteroid dehydrogenase 5 inhibitor) − 0.0526 × (Delayed rectifier potassium channel activator) − 0.0731 × (Cellulase inhibitor) + 0.0325 × (Insulin like growth factor 3 antagonist) + 0.0292 × (Leukotriene C antagonist) − 0.48 × (Volume^2) + 0.0212 × (Estrone sulfotransferase stimulant) + 0.0183 × (Skin irritation, high) − 0.0367 × (L-Threonine 3-dehydrogenase inhibitor) + 0.0312 × (4-Hydroxy-2-oxoglutarate aldolase inhibitor) + 0.0278 × (Toll-like receptor 4 agonist) + 0.00771 × (ATP-dependent RNA helicase inhibitor) + 0.0141 × (Tyrosine phenol-lyase inhibitor) − 0.0069 × (2-Amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase inhibitor) + 0.00643 × (Magnesium-protoporphyrin IX methyltransferase inhibitor) − 0.000826 × (Adenomatous polyposis treatment) − 0.000409 × (4-Hydroxymandelate oxidase inhibitor) + 0.729

Scheme 2.

Xplain GmbH) [20]. The X-Platform provides an integrated and comprehensive workflow management of a large number of "bricks", each providing specific function in the analysis of biological data. Figure 5 shows the relations between the selected targets. The analysis of the figure has revealed that the key elements of relationships of the selected targets are two proteins complexes GTP-binding proteins (G-alpha-s, G-beta, G-gamma and G-alpha-i, G-beta, G-gamma). These complexes are associated with the members of seven transmembrane domain superfamily of G-protein-coupled receptors that play the important role in the regulation of vital biological process in the organism.

3 Conclusions

In this study we found that our consensus approach has some benefits against single QSAR predictive models for acute rat toxicity with different routes of administrations. The proposed approach reveals comparable or higher accuracy of prediction, good coverage of the test sets and high performance. The validation on known set used for evaluation of T.E.S.T. 3.0 program is shown that the accuracy of the proposed approach considerably faster of them and has exceed the accuracy of both single and consensus methods represented by EPA. Moreover, this approach provides the proposals for biochemical and physiological mechanisms of acute toxicity, presenting a real example of system chemical toxicology.

Freely available on-line service for prediction of acute rat toxicity with four type of administration has been created based on the developed QSAR models:
http://www.pharmaexpert.ru/GUSAR/AcuToxPredict/

Acknowledgements

The study was partly supported by *ISTC* Grant # 3777, FP7 Grant # 200787 and Grant of the *Russian Ministry of Science and Education* (GK 02.740.11.0306).

Table 5. The most frequent activities from QSAR models on rat acute toxicity models for all type of administration.

Activity	Oral	IP[a]	IV[b]	SC[c]
3-Mercaptopyruvate sulfurtransferase inhibitor	4	8	6	5
Antibiotic Cephalosporin-like	8	7	7	5
Antineoplastic alkaloid	4	6	5	5
Antiosteoporotic	9	9	6	4
Aryldialkylphosphatase inhibitor	10	6	5	5
Carcinogenic, mouse, female	12	7	7	4
Carnitine dehydratase inhibitor	4	9	4	4
Complement C5a convertase stimulant	6	10	6	5
Glucose-6-phosphate isomerase inhibitor	9	14	4	4
Glyceryl-ether monooxygenase inhibitor	6	11	5	4
Hepatic disorders treatment	9	10	5	6
Leukotriene C4 antagonist	5	10	4	4
Leukotriene E4 antagonist	11	9	6	7
Nephrotoxic	8	11	4	8
Oxidizing agent	11	4	4	4
Prostaglandin EP2 agonist	8	5	5	5
Rotamase (FKBP12) inhibitor	4	7	5	5
Transforming growth factor antagonist	5	13	5	4
Acylglycerol lipase inhibitor	7	10	4	–
Adenosine A2b receptor agonist	1	5	6	6
12 Lipoxygenase inhibitor	8	10	5	–
Adenylate cyclase I inhibitor	9	9	–	5
2-Enoate reductase inhibitor	8	6	–	4
α_{1a}-Adrenoreceptor agonist	6	5	5	–
5 Hydroxytryptamine 1 antagonist	5	1	5	4
5 Hydroxytryptamine 1F agonist	5	4	6	3
5 Hydroxytryptamine 4A antagonist	8	4	9	1
5 Hydroxytryptamine 5 antagonist	–	7	6	7
Acetyl-CoA transferase inhibitor	8	11	4	–
Acetylglutamate kinase inhibitor	8	6	1	4

[a] Intraperitoneal route of administration; [b] Intravenous route of administration; [c] Subcutaneous route of administration.

References

[1] D. A. Filimonov, A. V. Zakharov, A. A. Lagunin, V. V. Poroikov, *SAR and QSAR Environ. Res.* **2009**, *20(7–8)*, 679 – 709.

[2] V. V. Poroikov, D. A. Filimonov, Yu.V. Borodina, A. A. Lagunin, A. Kos, *J. Chem. Inf. Comput. Sci.* **2000**, *40(6)*, 1349 – 1355.

[3] A. Sadym, A. Lagunin, D. Filimonov, V. Poroikov, *SAR QSAR Env. Res.* **2003**, *14*, 339 – 347.

[4] V. V. Poroikov, D. A. Filimonov, W.-D. Ihlenfeld, T. A. Gloriozova, A. A. Lagunin, Yu.V. Borodina, A. V. Stepanchikova, M. C. Nicklaus, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 228 – 236.

[5] A. A. Lagunin, A. V. Zakharov, D. A. Filimonov, V. V. Poroikov, *SAR QSAR Environ Res.* **2007**, *18(3–4)*, 285 – 298.

[6] E. Nielsen, G. Ostergaard, J. Ch. Larsen, *Toxicological Risk Assessment of Chemicals: A Practical Guide*, Informa HealthCare, New York **2008**, pp. 107 – 111.

[7] *SYMYX MDL Toxicity Database including RTECS*, Available at http://www.symyx.com/products/pdfs/toxds.pdf

[8] *ChemIDPlus*, available at http://chem.sis.nlm.nih.gov/chemidplus/

[9] *IUCLID*, available at http://iuclid.eu/index.php

[10] J. Devillers, H. Devillers, *SAR QSAR Environ Res.* **2009** *20(5–6)*, 467 – 500.

[11] A. Sazonovas, P. Japertas, R. Didziapetris, *SAR QSAR Environ Res.* **2010**, *21(1)*, 127 – 148.

[12] H. Zhu, T. M. Martin, L. Ye, A. Sedykh, D. M. Young, A. Tropsha, *Chem. Res. Toxicol.* **2009**, *22(12)*, 1913 – 1921.

[13] B. Lei, L. Xi, J. Li, H. Liu, X. Yao, *Anal. Chim. Acta.* **2009**, *644(1–2)*, 17 – 24.

[14] M. Hewitt, M. T. Cronin, J. C. Madden, P. H. Rowe, C. Johnson, A. Obi, S. J. Enoch, *J. Chem. Inf. Model.* **2007**, *47*, 1460 – 1468.

[15] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov, I. V. Tetko, *J. Chem. Inf. Model.* **2008**, *48(4)*, 766 – 784.

[16] J. S. Geman, E. Bienenstock, R. Doursat, *Neural Computation* **1992**, *4*, 1 – 58.

[17] *T.E.S.T.* 3.0, available at EPA web site: http://www.epa.gov/nrmrl/std/cppb/qsar/index.html.

[18] V. Poroikov, D. Filimonov, A. Lagunin, T. Gloriozova, A. Zakharov, *SAR QSAR Environ Res.* **2007**, *18(1–2)*, 101 – 110.

[19] S. M. Roberts, R. P. DeMott, R. C. James, *Drug Metab. Rev.* **1997**, *29(1–2)*, 329 – 353.

[20] *GeneXplain* web-site: http://www.genexplain.com

Received: October 31, 2010
Accepted: February 23, 2011
Published online: March 18, 2011