Acta Pharm. **62** (2012) 305–323
DOI: 10.2478/v10007-012-0024-z

Original research paper

Quantitative structure-pharmacokinetic relationship (QSPkR) analysis of the volume of distribution values of anti-infective agents from J group of the ATC classification in humans

BRUNO LOUIS[1]
VIJAY K. AGRAWAL[2]*

[1] *Department of Pharmacy, Sultan Qaboos University Hospital PO Box 38, Al Khod, Muscat 123 Oman*

[2] *QSAR and Computer Chemical Laboratories, A.P.S. University Rewa-486003, India*

Accepted June 12, 2012

In this study, a quantitative structure-pharmacokinetic relationship (QSPkR) model for the volume of distribution (V_d) values of 126 anti-infective drugs in humans was developed employing multiple linear regression (MLR), artificial neural network (ANN) and support vector regression (SVM) using theoretical molecular structural descriptors. A correlation-based feature selection (CFS) was employed to select the relevant descriptors for modeling. The model results show that the main factors governing V_d of anti-infective drugs are 3D molecular representations of atomic van der Waals volumes and Sanderson electronegativities, number of aliphatic and aromatic amino groups, number of beta-lactam rings and topological 2D shape of the molecule. Model predictivity was evaluated by external validation, using a variety of statistical tests and the SVM model demonstrated better performance compared to other models. The developed models can be used to predict the V_d values of anti-infective drugs.

Keywords: QSPkR, QSPR, structure pharmacokinetic relationship, volume of distribution, ANN, SVM, CFS

There is a constant need to develop new anti-infective drug molecules because the antibiotics and antiviral drugs always face a threat of resistance development, which may eventually lead to therapeutic failure. Drug discovery and development is a lengthy and costly process and many drug candidates fail at the later stages of development due to poor pharmacokinetic (PK) properties, *i.e.*, absorption, distribution, metabolism and excretion (ADME) (1). The success rate in drug discovery can be increased by predicting the PK properties of virtual molecules in the early stages and this will also reduce the cost of drug development. Computational or *in silico* methods are increasingly used in drug discovery to predict the properties of virtual molecules (2). Quantitative structure--activity relationships (QSARs) are mathematical models that attempt to relate the struc-

* Correspondence; e-mail: apsvka@yahoo.co.in; Present address: National Institute of Technical Teachers' Training and Research (NITTTR) Shamla Hills, Bhopal-462 002, India.

ture of a compound to its biological or physicochemical activity. Similarly, the quantitative structure-pharmacokinetic relationship (QSPkR) is used to model pharmacokinetic systems.

The volume of distribution (V_d) is an important PK property, which determines the extent of drug distribution in the body. The V_d, which is calculated as proportionality constant, considers drug distribution between organs and tissues to be homogeneous. It represents a measure of relative partitioning of the drug between plasma and the tissues. The V_d has a significant impact on other PK properties, such as clearance and half--life. Drugs having low V_d values require more frequent dosing intervals, whereas high values require less frequent dosing intervals. A higher V_d relates to greater tissue partitioning, which means that the drug can penetrate into tissues as well as bind reversibly to tissue components. For this reason, it is necessary for an administered drug to have appropriate tissue distribution. This is of particular interest to antibiotic dosing, where tissue distribution of the antibiotic mainly regulates the clinical effectiveness and toxicity. The tissue distribution reflected on V_d is also essential before making comparisons between antibiotics.

There are few studies on modeling the antibiotic V_d based on QSPkR and the attempts were generally confined to small sets of compounds and in some cases to sets of analogues. Chee *et al.* (3) developed QSPkR models to predict the volume of distribution of 44 antimicrobial agents in humans using the k-nearest-neighbor (k-NN) and partial least-square (PLS) methods. Turner *et al.* (4) reported an artificial neural network model to predict the volume of distribution and other PK properties for a series of 20 cephalosporins.

The present study was undertaken to establish a QSPkR model for predicting the volume of distribution of 126 anti-infective drugs belonging to different sub-therapeutic classes, antibacterial, antimycotic, antimycobacterial and antiviral, using theoretically calculated descriptors. In this attempt, we used nonlinear artificial neural network (ANN) and support vector machine (SVM) as well as multiple linear regression (MLR) methods. A large set of descriptors were calculated and a correlation based feature selection (CFS) method was employed to select the best set of descriptors for modeling.

EXPERIMENTAL

Dataset

To develop successful QSPkR models, reliable and good quality pharmacokinetic data is required. Human pharmacokinetic data is difficult to compile because PK values are derived from different studies in which experimental approaches differ. It is essential that data used in QSPkR modelling are obtained from studies in which the dose was administered intravenously. Thus, the V_d values are not confounded by effects of slow and incomplete absorption or extensive first-pass extraction. A carefully collected database of human pharmacokinetic parameters of drugs exclusively from intravenous administration was reported by Obach *et al.* (5). In this database, only intravenous data from rapid bolus injection or infusions were included and no data from oral, *i.m.*, or any other dosing routes. From this database, we have used V_d data of 126 anti-infective agents for

the present study. The selected drugs fall under the category of anti-infective (J) and subcategories antibacterial (J01), antimycotics (J02), antimycobacterial (J04) and antiviral (J05) according to the anatomical therapeutic classification (ATC) (6).

The volume of distribution values in steady state is expressed in liters per kilogram (L kg^{-1}). The V_d values range from 0.05 to 33 L kg^{-1} in the studied anti-infective drug dataset. The V_d values are converted to the logarithmic scale (log V_d) and then used as dependent variables like in the QSPkR analysis. A list of the anti-infective drugs and their corresponding log V_d values used in the present study is given in Supporting Information. External validation is an absolute requirement to obtain a truly predictive model (7). Therefore, the dataset was divided into a training set and test set for external validation. In this study, the division of the dataset into the training and test sets was done first by ranking the compounds according to their log V_d values and then taking every fifth compound as an external test compound and removing it from the dataset. The external test set of compounds was selected prior to the development of models.

Molecular descriptors

The molecular structure was searched using the PubChem Compound Database (8) and was built by using the ChemAxon software package (v.5.4, ChemAxon, Budapest, Hungary). All molecules were cleaned using standardizer tool available in ChemAxon to get uniform structure representations. The salts counter ions were removed and neutralized the atomic charges of molecules without producing valence errors on atoms. The molecular structure was then 3D optimized using CORINA (Molecular Networks, GmbH, Erlangen, Germany) and molecular descriptors were calculated using E-DRAGON (v.1, Milano Chemometrics, Milan, Italy); they are part of the on-line software provided by the Virtual Computational Chemistry Laboratory (9).

Descriptor selection and linear model generation

The selection of descriptors is an essential step in modeling. A large number of structural descriptors are generated by E-Dragon software. Therefore, the selection of proper descriptors to establish QSPkR models is a very important step to reduce over-fitting and improve the overall model predictability. Machine learning algorithms can be used for descriptor selection. In machine learning, first a descriptor sub-set is selected before the learning process. This reduces the dimensionality of data by removing unsuitable descriptors and improves the learning. In this study, we employed a filter based method called correlation-based feature selection (CFS) which was introduced by Hall and Holmes (10). CFS evaluates subsets of attributes rather than individual attributes. It is a subset evaluation heuristic that takes into account the usefulness of individual features for predicting the class along with the level of inter-correlation among them. The heuristic assigns high scores to subsets containing attributes that are highly correlated with the class and have low inter-correlation with each other:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (1)$$

where M_s is the »merit« of a feature subset S containing k features, $\overline{r_{cf}}$ is the average feature-class correlation and $\overline{r_{ff}}$ is the average feature-feature inter-correlation. The numerator of the above equation provides an indication of how predictive a set of features are; the denominator indicates how much redundancy there is among the features. The heuristic removes irrelevant features because they will be poor predictors of the class and redundant attributes that are highly correlated with one or more of the other features. In order to find the merit of a feature subset, it is necessary to compute correlation (dependence) between attributes. CFS first discretizes numeric features and then uses symmetrical uncertainty to estimate the degree of association between discrete features. After computing a correlation matrix, CFS applies a heuristic search strategy to find a good subset of features, according to the above equation.

We used the forward selection search, which produces a list of selected attributes. The reduced datasets was then passed to a machine learning scheme for linear modeling to produce a linear model. The entire process of descriptor selection and model building was implemented in WEKA data mining software (v.3.4, University of Waikato, Hamilton, New Zealand).

Validation techniques and model performance evaluation

The performance of the models was evaluated using the error measures, root mean square error (RMSE) and mean absolute error (MAE). They were calculated using the following formula:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\log V_{\mathrm{d\ obs}} - \log V_{\mathrm{d\ pred}}\right| \qquad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log V_{\mathrm{d\ obs}} - \log V_{\mathrm{d\ pred}})^2} \qquad (3)$$

where n is the number of data points, the parameter log $V_{\mathrm{d\ pred}}$ represents the predicted output from the model for a given input, while log $V_{\mathrm{d\ obs}}$ is the observed value for the same input.

We used a 10-fold cross-validation technique for selecting the models and their parameters. This procedure divides the dataset into 10 folds or groups and creates the model using 9 of the sets and tests it on the remaining group. This procedure is repeated until each of the 10 groups has served as a test group. The error estimates, RMSE and MAE are calculated and then averaged. Here, the 101 training dataset was randomly divided into 10 groups and the model was trained on 9 groups and the remaining group was used for testing each time.

The overall accuracy of predicted parameters was expressed in terms of average fold--error, which was calculated as the mean of the individual fold-error values (11). Fold error (FE) was calculated according to the following equation and the average values were reported as average fold-error AFE.

$$FE = \text{anti}\log\left|\log V_{\text{d obs}} - \log V_{\text{d pred}}\right| \quad (4)$$

Tropsha *et al*. (7) strongly advocate rigorous validation of models before their practical application or interpretation. To estimate the predictive power of a model, these authors recommended a set of statistical criteria: *(i)* correlation coefficient, *R*, between the predicted and observed activities; *(ii)* coefficients of determination (predicted *vs*. observed activities, R_0^2, and observed *vs.* predicted activities, $R_0'^2$, for regressions through the origin); *(iii)* slopes *k* and *k*' of regression lines (predicted *vs.* observed activities, and observed *vs.* predicted activities) through the origin. The model has an acceptable predictive power if the following conditions are satisfied:

$$R^2 > 0.6;\ \frac{R^2 - R_0^2}{R^2} < 0.1 \text{ and } 0.85 \le K \le 1.15$$

or

$$\frac{R^2 - R_0'^2}{R^2} < 0.1 \text{ and } 0.85 \le K' \le 1.15;\quad \left|R_0^2 - R_0'^2\right| < 0.3$$

Roy *et al.* (12) proposed another parameter, $R^2_{\text{m(test)}}$, for external predictability and this is calculated as:

$$R^2_{\text{m(test)}} = R^2\left(1 - \sqrt{R^2 - R_0^2}\right) \quad (5)$$

where R_0^2 is the squared correlation coefficient between the observed and predicted values of the test set compounds with intercept set to zero. The recommended value of $R^2_{\text{m(test)}}$ for successful predictability is greater than 0.5.

Relevant descriptors for modeling were selected from a large pool of 1664 Dragon descriptors. It is necessary to check the possible existence of any chance correlation. A randomization or permutation test was carried out to check the existence of chance correlation. The dependent variable, log V_{d}, was randomly shuffled and a new QSPkR model was developed, using the MLR algorithm. The procedure was repeated several times and the new models are expected to have low R^2 values. This was further tested by a new parameter, R_{p}^2, which penalizes the model R^2 for the difference between the squared mean correlation coefficient (R_{r}^2) of randomized models and squared correlation coefficient (R^2) of the non-randomized model (12). The R_{p}^2 is calculated by the following equation:

$$R_{\text{p}}^2 = R^2\sqrt{R^2 - R_{\text{r}}^2} \quad (6)$$

The value of R_{p}^2 should be greater than 0.5 for an acceptable model to ensure that the developed model was not obtained by chance.

Artificial neural network

The theory and application of ANN studies in QSPR modelling is extensively discussed in many reviews (13). A conventional three-layered back-propagation network was employed in this study. The back-propagation ANN uses the supervised learning technique and the network is trained by minimizing the squared error of the network's output. The error is calculated between the desired values and the network's output. This error is propagated backwards through the network for adjusting the weights to minimize the error.

The architecture of the network consists of five neurons in the input layer, which correspond to the five significant descriptors selected in the linear model and one neuron in the output layer which is the log V_d value. A sigmoid transfer function was used for each neuron in all layers. The number of neurons in the hidden layer was selected by varying the number and the best model giving the lowest RMSE and the highest correlation coefficient. The learning rate and momentum parameters were optimized by varying the values and the optimum was selected. The over-fitting problem was minimized by monitoring the performance of the network during training by using a validation dataset. The software has a facility to select a certain percentage of data as a validation set and we used 20 % of the training dataset for validation. A 10-fold cross validation was done during the selection of the above network parameters.

Support vector machines

A support vector machine (SVM) is a supervised learning algorithm originally developed for pattern classification problems. This technique has had much success in QSAR modeling studies and the details are given in literature (14). In SVM, the input data is first mapped into a high-dimensional feature space by the use of kernel function and then linear regression is performed in the feature space. The nonlinear feature mapping will allow the treatment of nonlinear problems in a linear space. In the higher dimensional feature space, SVM approximates the set of data with a linear function:

$$y = \sum_{i=1}^{m} w_i \Phi(x_i) + b \tag{7}$$

where $\Phi(x_i)$ are the features of input variables after kernel transformation while w_i and b are coefficients. The radial basis function (RBF) kernel is commonly used in QSPR problems. The RBF kernel can perform nonlinear mapping as described by the following equation:

$$k(x,y) = \exp(-\gamma \|x - y\|^2) \tag{8}$$

After kernel transformation, the new feature space allows the data to be linearly separable by hyper planes or conduct a linear regression. Coefficient w and b are estimated by minimizing the regularized risk function which is defined as:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L\varepsilon(y_i, f(x_i, w)) + \frac{1}{2}\|w\|^2 \qquad (9)$$

The first term $C\frac{1}{N}\sum_{i=1}^{N} L\varepsilon(y_i, f(x_i, w))$ is the empirical error (risk) and it is measured by ε insensitive loss function, where ε is a prescribed parameter and is referred to as the tube size, and it is defined as the approximation accuracy placed on the training data points. The loss function ignores errors as long as it is less than ε; in other words, errors below ε would not be penalized. The second term $\frac{1}{2}\|w\|^2$, the regularization term, is a measure of function flatness. The value of the cost function C determines the regularized constant and determines the trade off between the empirical error and the regularized term. Minimization of the regularized risk function is a constrained optimization problem that can be reformulated into dual problem formalism by using Lagrange multipliers. The support vector regression calculations are performed using John Platt's sequential minimal optimization (SMO) algorithm modified by Smola and Scholkopf in WEKA software (15).

The first step in the SVM model construction is selection of the kernel type and then optimization of the kernel parameter. We selected the RBF kernel, so that the value of kernel parameter γ (gamma) would be optimized. The next step was optimizing parameter ε of ε-insensitive loss function, and complexity parameter or regularization parameter C. After optimizing the values C, ε and γ the support vector machine was first trained on a training dataset having known log V_d values and the trained SVM was used to predict log V_d values for test data. SVM performance depends on selection of the kernel type and optimizing parameters C, ε and γ, so a ten-fold cross validation is used in the optimization process.

RESULTS AND DISCUSSION

Descriptor selection and linear model

By using E-DRAGON, a total of 1664 descriptors were computed, including 1D, 2D and 3D descriptors. The V_d data of 126 anti-infective drugs was partitioned into an external test set of 25 compounds and a training set of 101 compounds. The selection of relevant descriptors is an important step for constructing a predictive model. The training dataset was used to describe selection using the CFS method. This method estimated a subset of 20 descriptors from a pool of 1664 dragon descriptors. The reduced descriptor subset was then used for linear model building using a forward step-wise multiple linear regression analysis. The resulting linear model selected five descriptors to give a stable model with R = 0.860, RMSE = 0.239, MAE = 0.181. The selected descriptor variables are nNR2, PJI2, Mor23v, Mor28e and nBlct and their values are given in Supporting Information.

The resulting equation with five descriptors is as follows:

$$\log V_d = 0.269(\pm 0.456) + 0.205(\pm 0.071)\ \text{nNR2} - 0.682(\pm 0.503)\ \text{PJI2} - 0.512(\pm 0.275)\ \text{Mor23v} - 0.193(\pm 0.127)\ \text{Mor28e} - 0.433(\pm 0.109)\ \text{nBlct}$$

$$N = 101,\ R = 0.860\ R^2 = 0.740 \qquad \text{SE} = 0.247\ R^2_{cv} = 0.648 \qquad (10)$$

where R is the correlation coefficient, SE is the standard error of estimate. The figures in parentheses with the regression coefficients are standard errors of coefficients. The correlation matrix for selected parameters nNR2, PJI2, Mor23v, Mor28e and nBlct and also for log V_d is given in Table I. The correlation matrix shows no intercorrelation of selected descriptors. This was further confirmed by calculating the variance inflation factor (VIF) and tolerance. The VIF = $1/1–R^2$ and tolerance = 1/VIF. In practice, when VIF > 5 or if the tolerance remains less than 0.20, then this would indicate multicollinearity amongst the descriptors. The calculated VIF and tolerance values are given in Table II. There is no multicollinearity problem for the selected five descriptors. The developed model was cross-validated by the leave-one-out (LOO) method and the calculated cross-validation parameter R^2_{cv} (cross validated correlation coefficient). The high value observed is indicative of their reliability in prediction.

Table I. Inter-correlation of descriptors

	log V_d	nNR2	PJI2	Mor23v	Mor28e	nBlct
log V_d	1.000					
nNR2	0.665	1.000				
PJI2	–0.242	–0.022	1.000			
Mor23v	–0.446	–0.309	0.003	1.000		
Mor28e	–0.446	–0.348	0.204	0.402	1.000	
nBlct	–0.637	–0.388	0.110	0.100	0.046	1.000

nNR2 – total number of aromatic and aliphatic tertiary amino groups, PJI2 – Petitjean shape index, Mor23v – 3D-MoRSE signal 23/weighted by atomic van der Waals volumes, Mor28e-3D-MoRSE signal 28/weighted by atomic Sanderson electronegativities, nBlct – number of beta-lactam rings

Table II. Collinearity statistics

Descriptor	Tolerance	VIF
nNR2	0.712	1.405
PJI2	0.932	1.074
Mor23v	0.800	1.250
Mor28e	0.731	1.368
nBlct	0.825	1.212

VIF – variance inflation factor

Table III. Results of randomization test

Iteration	R^2	Iteration	R^2
1	0.051	11	0.020
2	0.019	12	0.040
3	0.028	13	0.026
4	0.024	14	0.058
5	0.101	15	0.022
6	0.052	16	0.035
7	0.095	17	0.045
8	0.076	18	0.041
9	0.013	19	0.042
10	0.033	20	0.051

The obtained MLR model was validated using the *Y*-randomization tests to determine the probability of chance correlation during descriptor selection. Twenty random shuffles of the log V_d values were chosen, and the models were developed for the training dataset, using the original descriptor matrix. All the models obtained in the randomization test have very low R^2 values, which are shown in Table III. Lower values of R^2 in comparison with the real model's results corroborate that the descriptor selection by the

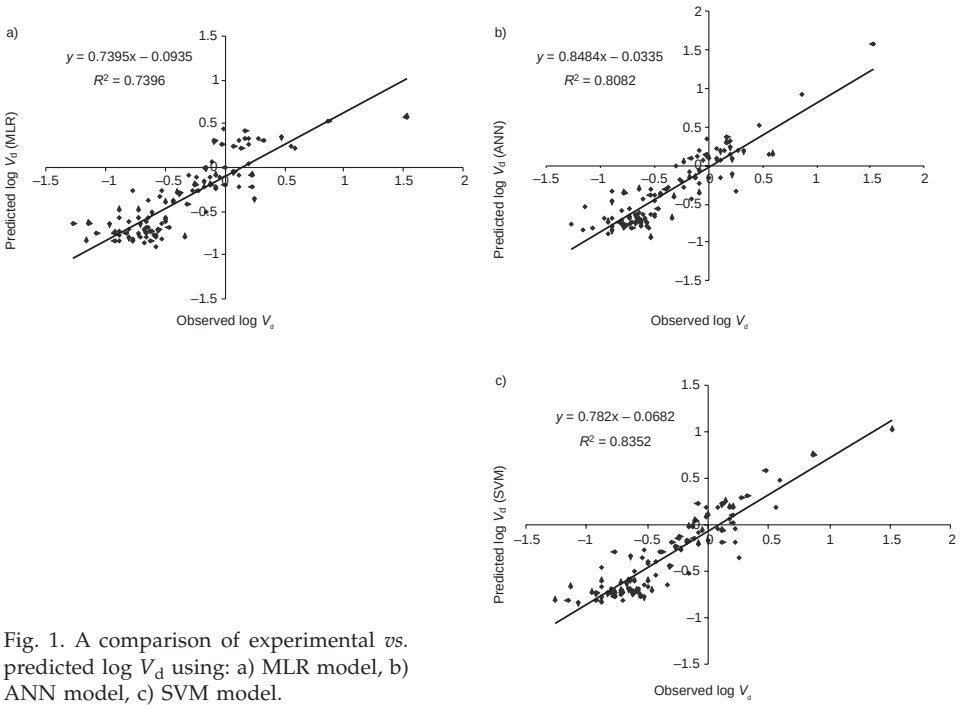

Fig. 1. A comparison of experimental *vs.* predicted log V_d using: a) MLR model, b) ANN model, c) SVM model.

MLR model was not due to a chance correlation. There is a real relationship between the molecular structure description (descriptors) and the log V_d values of the studied set. The calculated value of R_p^2 for the model is 0.617 and this is above the stipulated value of 0.5, which further proves that the model developed was not by chance.

The predicted value of log V_d of the training dataset, using this model, is plotted against experimental values and is shown in Fig. 1a. The above linear model was used to predict the 25 external test data set which was never used in descriptor selection or model building. The result shows R^2 = 0.671, MAE = 0.197 and RMSE = 0.280. The predicted values of log V_d of the training and test set using the MLR equation are given in Table IV.

The selected five descriptors, *i.e.*, nNR2, PJI2, Mor23v, Mor28e and nBlct, provide some insight into the structural influence on the volume of distribution of anti-infective agents in humans.

The 3D-MoRSE descriptors Mor23v and Mor28e are the 3D-MoRSE-signal 23/weighted by atomic van der Waals volumes (Mor23v) and 3D-MoRSE-signal 28/weighted by atomic Sanderson electronegativities (Mor28e) (16). Petitjean shape index (PJI2) is a topological 2D descriptor (17). The other two descriptors, nNR2 and nBlct, represent the functional group count in the molecule. The nNR2 is the total number of aromatic (nArNR2) and aliphatic ($nRNR_2$) tertiary amino groups present in the molecule while nBlct is the number of beta-lactam rings in the molecule.

The 3D-MoRSE descriptors are the 3D molecular representations of the structure based on electron diffraction; they are calculated by summing up atomic mass viewed by a different angular scattering function (16). The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle(s) in the range of 0–31A$^{0-1}$ from the three-dimensional atomic coordinates of a molecule. The 3D-MoRSE descriptor is calculated using the following expression:

$$Morsw = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (11)$$

w_i and w_j are the characteristic properties of atoms i and j (including unweighted masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities), r_{ij} is the interatomic distance, and nAT is the number of atoms in the molecule, and s is the scattering angle. The selected descriptors Mor23v and Mor28e in the linear model reflect the negative influence of van der Waals volumes and electronegativities on the V_d values.

Petitjean shape index (PJI2) is a topological anisometry descriptor (17). It is also called graph-theoretical shape coefficient and PJI2 and is defined as:

$$PJI2 = \frac{D - R_{ad}}{R_{ad}} \quad 0 \le I_2 \le 1 \qquad (12)$$

where R_{ad} is the topological radius and D is the topological diameter obtained from the distance matrix representing the molecular graph. The shape of a molecule is an important property determining many biological properties; here we observe that the topological shape of the molecule has a negative influence on log V_d values.

Table IV. Observed and predicted values of log V_d and the residual values

Drug	obs log V_d	MLR predict	Residual	ANN predict	Residual	SVM predict	Residual
Abacavir	–0.076	–0.077	–0.001	–0.076	0	–0.074	0.002
Adefovir	–0.377	–0.297	0.080	–0.294	0.083	–0.290	0.087
Amdinocillin	–0.432	–0.378	0.054	–0.384	0.048	–0.545	–0.113
Amikacin	–0.796	–0.696	0.100	–0.588	0.208	–0.683	0.113
Amoxicillin	–0.602	–0.728	–0.126	–0.645	–0.043	–0.593	0.009
Ampicillin	–0.658	–0.796	–0.138	–0.741	–0.083	–0.683	–0.025
Azithromycin	1.519	0.583	–0.936	1.580	0.061	1.029	–0.490
Biapenem	–0.699	–0.643	0.056	–0.746	–0.047	–0.719	–0.020
Cefamandole	–0.796	–0.803	–0.007	–0.761	0.035	–0.724	0.072
Cefatrizine	–0.658	–0.745	–0.087	–0.658	0	–0.606	0.052
Cefazolin	–0.921	–0.848	0.073	–0.895	0.026	–0.811	0.110
Cefcanel	–0.886	–0.775	0.111	–0.691	0.195	–0.661	0.225
Cefepime	–0.553	–0.712	–0.159	–0.812	–0.259	–0.758	–0.205
Cefixime	–0.620	–0.843	–0.223	–0.784	–0.164	–0.729	–0.109
Cefodizime	–1.268	–0.635	0.633	–0.764	0.504	–0.802	0.466
Cefoperazone	–0.770	–0.849	–0.079	–0.804	–0.034	–0.761	0.009
Ceforanide	–0.770	–0.848	–0.078	–0.778	–0.008	–0.712	0.058
Cefotaxime	–0.721	–0.759	–0.038	–0.763	–0.042	–0.749	–0.028
Cefotetan	–0.886	–0.735	0.151	–0.803	0.083	–0.774	0.112
Cefoxitin	–0.770	–0.813	–0.043	–0.785	–0.015	–0.762	0.008
Cefprozil	–0.678	–0.712	–0.034	–0.656	0.022	–0.629	0.049
Ceftizoxime	–0.699	–0.790	–0.091	–0.812	–0.113	–0.812	–0.113
Ceftobiprole	–0.569	–0.901	–0.332	–0.795	–0.226	–0.682	–0.113
Ceftriaxone	–1.071	–0.746	0.325	–0.817	0.254	–0.835	0.236
Cephalexin	–0.678	–0.861	–0.183	–0.756	–0.078	–0.629	0.049
Cephaloridine	–0.337	–0.787	–0.450	–0.682	–0.345	–0.634	–0.297
Cephalothin	–1.155	–0.830	0.325	–0.845	0.310	–0.814	0.341
Cephapirin	–0.886	–0.837	0.049	–0.857	0.029	–0.815	0.071
Chlortetracycline	–0.046	–0.109	–0.063	0.066	0.112	–0.056	–0.010
Cidofovir	–0.310	–0.415	–0.105	–0.405	–0.095	–0.443	–0.133
Ciprofloxacin	0.322	0.311	–0.011	0.180	–0.142	0.309	–0.013
Clarithromycin	0.176	0.419	0.243	0.368	0.192	0.194	0.018
Clavulanic Acd	–0.658	–0.688	–0.030	–0.710	–0.052	–0.574	0.084
Clinafloxacin	0.279	0.325	0.046	0.186	–0.093	0.293	0.014
Clindamycin	–0.102	0.073	0.175	0.105	0.207	0.052	0.154
Dapsone	–0.081	–0.227	–0.146	–0.357	–0.276	–0.194	–0.113

Demethylchlor-tetracycline	0.114	–0.086	–0.200	0.070	–0.044	–0.054	–0.168
Dibekacin	–0.886	–0.580	0.306	–0.478	0.408	–0.598	0.288
Dicloxacillin	–0.959	–0.650	0.309	–0.696	0.263	–0.714	0.245
Didanosine	–0.114	–0.186	–0.072	–0.163	–0.049	–0.150	–0.036
Doxycycline	–0.161	–0.012	0.149	0.093	0.254	–0.008	0.153
Ertapenem	–0.921	–0.740	0.181	–0.737	0.184	–0.734	0.187
Erythromycin	–0.022	0.268	0.290	0.134	0.156	0.186	0.208
Ethambutol	0.230	–0.217	–0.447	–0.126	–0.356	–0.192	–0.422
Fleroxacin	0.204	0.330	0.126	0.309	0.105	0.101	–0.103
Flucloxacillin	–0.721	–0.580	0.141	–0.662	0.059	–0.698	0.023
Fluconazole	–0.125	–0.102	0.023	–0.150	–0.025	–0.011	0.114
Flucytosine	–0.167	–0.167	0	–0.090	0.077	–0.172	–0.005
Foscarnet	–0.301	–0.084	0.217	–0.011	0.290	–0.188	0.113
Fosfomycin	–0.495	–0.392	0.103	–0.379	0.116	–0.412	0.083
Ganciclovir	0	–0.202	–0.202	–0.166	–0.166	–0.170	–0.170
Imipenem	–0.620	–0.722	–0.102	–0.756	–0.136	–0.707	–0.087
Indinavir	–0.086	0.299	0.385	0.191	0.277	0.231	0.317
Isepamicin	–0.495	–0.584	–0.089	–0.463	0.032	–0.594	–0.099
Itraconazole	0.869	0.538	–0.331	0.927	0.058	0.758	–0.111
Kanamycin	–0.585	–0.764	–0.179	–0.708	–0.123	–0.699	–0.114
Lamivudine	0.114	–0.215	–0.329	–0.158	–0.272	–0.190	–0.304
Levofloxacin	0.079	0.250	0.171	0.211	0.132	0.191	0.112
Lincomycin	0	–0.001	–0.001	0.095	0.095	0.113	0.113
Linezolid	–0.237	–0.108	0.129	0.050	0.287	–0.123	0.114
Meropenem	–0.523	–0.816	–0.293	–0.941	–0.418	–0.774	–0.251
Methicillin	–0.495	–0.619	–0.124	–0.663	–0.168	–0.662	–0.167
Metronidazole	–0.398	–0.275	0.123	–0.312	0.086	–0.285	0.113
Minocycline	0.204	0.051	–0.153	0.140	–0.064	0.030	–0.174
Moxifloxacin	0.146	0.228	0.082	0.192	0.046	0.256	0.110
Nafcillin	–0.658	–0.614	0.044	–0.661	–0.003	–0.688	–0.030
Netilmicin	–1.137	–0.626	0.511	–0.530	0.607	–0.656	0.481
Nitrofurantoin	–0.244	–0.192	0.052	–0.213	0.031	–0.147	0.097
Ofloxacin	0.204	0.265	0.061	0.226	0.022	0.197	–0.007
Oseltamivir acid	–0.432	–0.386	0.046	–0.361	0.071	–0.396	0.036
Oxacillin	–0.721	–0.581	0.140	–0.685	0.036	–0.727	–0.006
Oxytetracycline	0.230	–0.079	–0.309	0.079	–0.151	–0.036	–0.266
Panipenem	–0.721	–0.476	0.245	–0.579	0.142	–0.608	0.113
Pefloxacin	0.176	0.330	0.154	0.300	0.124	0.065	–0.111
Penicillin G	–0.620	–0.726	–0.106	–0.729	–0.109	–0.730	–0.110

Phenethicillin	–0.523	–0.641	–0.118	–0.642	–0.119	–0.634	–0.111
Piperacillin	–0.569	–0.786	–0.217	–0.752	–0.183	–0.721	–0.152
Ribostamycin	–0.602	–0.514	0.088	–0.445	0.157	–0.490	0.112
Rifampin	–0.013	0.439	0.452	0.336	0.349	0.097	0.11
Saquinavir	0.556	0.239	–0.317	0.144	–0.412	0.189	–0.367
Sparfloxacin	0.591	0.221	–0.370	0.150	–0.441	0.481	–0.110
Spectinomycin	–0.886	–0.470	0.416	–0.347	0.539	–0.461	0.425
Stavudine	–0.174	–0.216	–0.042	–0.156	0.018	–0.191	–0.017
Streptomycin	–0.469	–0.687	–0.218	–0.567	–0.098	–0.696	–0.227
Sulbactam	–0.495	–0.579	–0.084	–0.648	–0.153	–0.395	0.100
Sulbenicillin	–0.824	–0.754	0.070	–0.743	0.081	–0.734	0.090
Sulfadiazine	–0.538	–0.325	0.213	–0.379	0.159	–0.354	0.184
Sulfamethoxazole	–0.523	–0.257	0.266	–0.311	0.212	–0.266	0.257
Sulfisoxazole	–0.770	–0.270	0.500	–0.351	0.419	–0.282	0.488
Telithromycin	0.477	0.345	–0.132	0.513	0.036	0.588	0.111
Tenofovir	–0.081	–0.234	–0.153	–0.239	–0.158	–0.199	–0.118
Tetracycline	0.079	–0.049	–0.128	0.084	0.005	–0.035	–0.114
Ticarcillin	–0.796	–0.614	0.182	–0.715	0.081	–0.713	0.083
Tinidazole	–0.229	–0.259	–0.030	–0.286	–0.057	–0.258	–0.029
Tobramycin	–0.638	–0.690	–0.052	–0.622	0.016	–0.700	–0.062
Tomopenem	–0.638	–0.581	0.057	–0.708	–0.070	–0.589	0.049
Trospectomycin	–0.155	–0.496	–0.341	–0.448	–0.293	–0.519	–0.364
Trovafloxacin	0.114	0.312	0.198	0.183	0.069	0.228	0.114
Zalcitabine	–0.268	–0.259	0.009	–0.186	0.082	–0.231	0.037
Zanamivir	–0.638	–0.376	0.262	–0.303	0.335	–0.336	0.302
Zidovudine	0.255	–0.364	–0.619	–0.332	–0.587	–0.357	–0.612
Test data							
Acyclovir	–0.149	–0.203	–0.054	–0.168	–0.019	–0.170	–0.021
Azlocillin	–0.585	–0.599	–0.014	–0.664	–0.079	–0.692	–0.107
Aztreonam	–0.745	–0.775	–0.030	–0.788	–0.043	–0.795	–0.050
Carbenicillin	–0.770	–0.742	0.028	–0.761	0.009	–0.779	–0.009
Cefadroxil	–0.638	–0.796	–0.158	–0.651	–0.013	–0.551	0.087
Cefetamet	–0.553	–0.805	–0.252	–0.817	–0.264	–0.808	–0.255
Cefmetazole	–0.886	–0.847	0.039	–0.947	–0.061	–0.768	0.118
Cefpirome	–0.620	–0.754	–0.134	–0.773	–0.153	–0.789	–0.169
Ceftazidime	–0.509	–0.759	–0.250	–0.749	–0.240	–0.751	–0.242
Cefuroxime	–0.824	–0.770	0.054	–0.798	0.026	–0.81	0.014
Cephradine	–0.678	–0.843	–0.165	–0.866	–0.188	–0.815	–0.137
Chloramphenicol	–0.027	–0.257	–0.230	–0.230	–0.203	–0.219	–0.192
Gatifloxacin	0.230	0.226	–0.004	0.150	–0.080	0.496	0.266

Genaconazole	–0.208	–0.178	0.030	–0.166	0.042	–0.135	0.073
Gentamicin	–0.481	–0.231	0.250	–0.126	0.355	–0.132	0.349
Isoniazide	–0.086	–0.366	–0.280	–0.378	–0.292	–0.393	–0.307
Mezlocillin	–1.046	–0.550	0.496	–0.671	0.375	–0.746	0.300
Nevirapine	0.114	0.290	0.176	0.142	0.028	0.082	–0.032
Penciclovir	0.041	–0.102	–0.143	–0.087	–0.128	–0.101	–0.142
Phenoxymethylpenicillin	–0.387	–0.603	–0.216	–0.640	–0.253	–0.648	–0.261
Rolitetracycline	–0.268	0.146	0.414	0.149	0.417	0.177	0.445
Sisomicin	–0.721	–0.663	0.058	–0.580	0.141	–0.690	0.031
Sitafloxacin	0.176	0.201	0.025	0.147	–0.029	0.432	0.256
Tigecycline	1.079	0.231	–0.848	0.200	–0.879	0.323	–0.756
Voriconazole	0.342	–0.232	–0.574	–0.273	–0.615	–0.226	–0.568

The total number of aromatic (nArNR2) and aliphatic (nRNR$_2$) amino groups present in the molecule has a positive influence on log V_d values of anti-infective drugs. Incorporation of a basic centre in a molecule will increase the volume of distribution. This is because the basic drugs are ionized at physiological pH due to ion-pair interactions between the basic centre and the charged acidic head groups of phospholipid membranes resulting in high affinity. An example is the introduction of a second basic centre into the macrolide aglycone ring in azithromycin, which results in a dramatic increase in the volume of distribution of 33 L kg^{-1} (log V_d = 1.519) compared to the macrolide antibiotic erythromycin with 0.95 L kg^{-1} (log V_d = –0.022) (Fig. 2).

Finally, the number of beta-lactam rings present in the molecules is also an important variable determining the V_d values of anti-infective. This specific molecular feature indicates the low volume of distribution of compounds containing such structures.

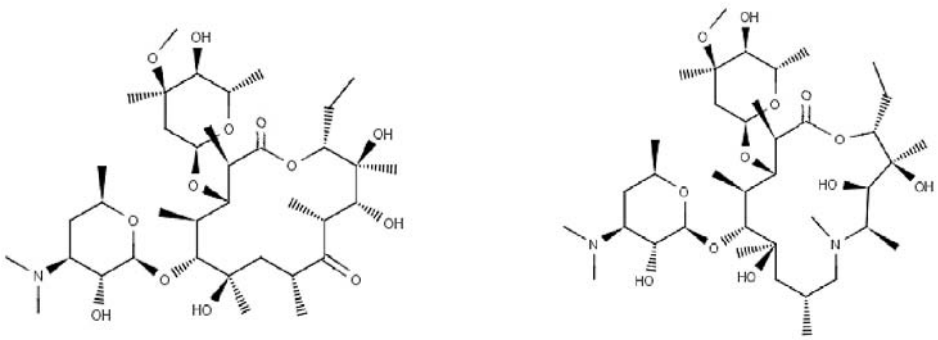

Fig. 2. Structures of antibiotics erythromycin (nNR2 = 1) and azithromycin (nNR2 = 2).

ANN models

The initial architecture of ANN was five neurons in the input layer and three neurons in the hidden layer selected by the auto build function and one output neuron. Input neurons correspond to the five selected descriptors nNR2, PJI2, Mor23v, Mor28e and nBlct. A sigmoid transfer function was used in all layers. The ideal value of learning rate η and momentum μ was determined by varying their values from 0.01 to 1.0 and the combination of $\eta = 0.21$ and $\mu = 0.61$, which gives the lowest RMSE, was selected. Optimization was done with 10-fold cross validation and 20 % of data used for validation. The learning time or number of epoch selected was 628. With the above selected parameters, the number of neurons in the hidden layer was optimized by varying from 1 to 10 and the ANN model with three hidden neurons gave the best performance. When the entire training data was trained in the network with the architecture of 5-3-1 and optimized parameters, it gave $R^2 = 0.808$, RMSE = 0.207 and MAE = 0.155. The plot of the experimental and predicted value of log V_d of the training data using the ANN model is shown in Fig. 1b.

Using the trained network, the test set was used for prediction and gave $R^2 = 0.671$, RMSE = 0.284 and MAE = 0.197. The predicted values of log V_d of the training and test data are given in Table IV.

SVM models

Optimization of SVM parameters was performed by systemically varying the parameter values in the training step using 10-fold cross validation and calculating the RMSE of the model. The parameter value that gave the lowest RMSE was selected. To make the learning process stable, a high value should be initially set up for *C*. We initially kept the value of *C* as 100 and optimized kernel parameter γ and tube size ε. The regularization parameter *C* controls the trade off between maximizing the margin and minimizing the training error. If the value of *C* is too low, then insufficient stress will be placed on the fitting of training data.

The RBF kernel parameter γ controls the amplitude of the Gaussian function and further affects the generalization ability of SVM. To obtain the optimal γ, the support vector learning machines were trained with γ values varying from 0.01 to 0.5. The optimum value was selected as 0.3, which gave the lowest RMSE. Parameter ε of ε-insensitive loss function is referred to as the tube size and is defined as the approximation accuracy placed on the training data points. The value of ε also determines the number of support vectors. The higher the value, the fewer support vectors are selected.

The optimum value of ε was found by varying the value 0.01 to 0.2 and the value 0.04 gave the lowest RMSE. After finding the values of ε and γ, the C value was further optimized as 105. The selected parameters ($\gamma = 0.3$, $\varepsilon = 0.04$, $C = 105$) were used for the final training run on the training set and predicted the log V_d values. The plot of predicted *vs.* experimental log V_d based on this model is shown in Fig. 1c and the values are shown in Table IV. The statistical parameters of this model are RMSE = 0.191, $R^2 = 0.835$ and MAE = 0.145 for the training set.

This SVM model is used to predict log V_d values of the test data set and the values are given in Table IV. The prediction statistics is RMSE = 0.273, $R^2 = 0.683$, and MAE = 0.207.

Comparison of MLR, ANN and SVM models

A summary of the performance of these three models is shown in Table V. Judging by R^2, RMSE and MAE, the SVM method gave the best performance for the training data set used in the present study. This model gave the highest R^2 and lowest error compared to other models. The training set prediction involved 1.46 average fold errors. A predicted method with an AFE ≤ 2 (*i.e.*, between 0.5 and 2.0) was considered successful for pharmacokinetic parameters (18). For the SVM model, the percentage of compounds with 2-fold error was 87 compared to 86 and 79 for ANN and MLR models. These statistic results indicate the good predictive power of this method, especially considering the error margin for pharmacokinetic data.

Analysis of the training set fit results would suggest that the SVM models will predict log V_d more effectively than MLR and ANN models on account of their higher R^2 and lower error values. However, this can only be proven after conducting rigorous cross validation procedures.

The model predictivity was assessed by five different statistical parameters for the external test set of 25 compounds and the results are shown in Table VI. All the models

Table V. Comparison between MLR, ANN and SVM models using the training dataset[a]

Statistics	MLR	ANN	SVM
R^2	0.740	0.808	0.835
RMSE	0.239	0.207	0.191
MAE	0.181	0.155	0.145
AFE	1.648	1.514	1.466

[a] N = 101.

Table VI. Predictivity of the external test set[a]

Statistics	MLR	ANN	SVM
R^2	0.671	0.671	0.683
$R^2 - R_0^2/R^2$	0.017	0.028	0.002
$R^2 - R_0'^2/R^2$	0.220	0.280	0.103
$\lvert R_0^2 - R_0'^2 \rvert$	0.136	0.169	0.069
K	0.910	0.899	0.895
K'	0.846	0.851	0.877
$R^2_{m(test)}$	0.599	0.579	0.659
AFE	1.80	1.82	1.785
RMSE	0.280	0.284	0.273
MAE	0.197	0.197	0.207

[a] N = 25.

show good and acceptable external predictive power and satisfy the following conditions for the external test set: $R^2 > 5$, $R^2 - R_0^2/R^2$ or $R^2 - R_0'^2/R^2 =< 0.1$, and K or K' within 0.85–1.15, $\left|R_0^2 - R_0'^2\right| < 0.3$, and $R^2_{\mathrm{m(test)}} > 0.5$.

Examination of the results of RMSE, AFE and MAE values in prediction of the independent test set shows that all the models have similar predictive power.

The AFE of the test set is less than 2 for all models; it is 1.78 for the SVM model, while 1.82 and 1.80 for ANN and MLR models, respectively.

CONCLUSIONS

Prediction of the volume of distribution of new chemical entities in humans is important in drug discovery and development. Use of predictive QSPkR modeling approaches may allow one to select drug candidates with desired pharmacokinetic properties. The V_d is a key pharmacokinetic parameter in determining the dosing regimen. Anti-infective agents from J group of ATC classification need to demonstrate adequate pharmacokinetic behaviour, permitting convenient dosing regimens that result in high patient compliance and thus effective therapy. We have demonstrated the feasibility of constructing QSPkR models using MLR, ANN and SVM methods for prediction of human V_d of 126 anti-infective agents with diverse chemical structures. We developed models and estimated the V_d values using five descriptors selected using the CFS method from a large pool of descriptors. The present study identified and provided some key chemical structural factors that effect V_d of anti-infective agents. The number of aromatic and aliphatic tertiary amino groups present in the molecule increases the V_d values of anti--infective drugs. The van der Waals volume and Sanderson electronegativities representing the 3D Morse descriptor, 2D shape of the molecules and the presence of beta-lactam ring system have a negative influence on the volume of distribution. Models proposed here satisfy all the rigorous criteria adopted for validation. All of the developed models show enhanced prediction capability. The results indicate that the SVM model could be useful for predicting the log V_d value.

Supporting information are available in electronic version of the article.

REFERENCES

1. T. Kennedy, Managing the drug discovery and development interface, *Drug Discov. Today* **2** (1997) 436–444; DOI: 10.1016/S1359-6446(97)01099-4.
2. M. Brvar, A. Perdih, V. Hodnik, M. Renko, G. Anderluh, R. Jerala and T. Solmajer, In silico discovery and biophysical evaluation of novel 5-(2-hydroxybenzylidene) rhodanine inhibitors of DNA gyrase B, *Bioorg. Med. Chem.* **20** (2012) 2572–2580; DOI: 10.1016/j.bmc.2012.02.052.
3. N. G. Chee, Y. Xiao, W. Putnam, B. Lum and A. Tropsha, Quantitative structure-pharmacokinetic parameters relationships (QSPKR) analysis of antimicrobial agents in humans using simulated annealing k-nearest-neighbor and partial least-square analysis methods, *J. Pharm. Sci.* **93** (2004) 2535–2544; DOI: 10.1002/jps.20117.

4. J. V. Turner, D. J. Maddalena, D. J. Cutler and S. Agatonovic-Kustrin, Multiple pharmacokinetic parameter prediction for a series of cephalosporins, *J. Pharm. Sci.* **92** (2003) 552–559; DOI: 10.1002/jps.10314.
5. R. S. Obach, F. Lombardo and N. J. Waters, Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 compounds, *Drug Metab. Dispos.* **36** (2008) 1385–1405; DOI: 10.1124/dmd.108.020479.
6. World Health Organization Collaborating Centre for Drug Statistics Methodology, *Guidelines for ATC classification and DDD assignment 2010*, 13th ed., WHO Collaborating Centre for Drug Statistics Methodology, Oslo 2009, pp. 163–177.
7. A. Tropsha, P. Gramatica and V. K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.* **22** (2003) 69–77; DOI: 10.1002/qsar.200390007.
8. National Center for Biotechnology Information, PubChem Compound Database; http://pubchem.ncbi.nlm.nih.gov/access January 20, 2011.
9. I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk and V. V. J. Prokopenko, Virtual computational chemistry laboratory – design and description, *Comput. Aid. Mol. Des.* **19** (2005) 453–463; DOI: 10.1007/s10822-005-8694-y.
10. M. A. Hall and G. Holmes, Benchmarking attribute selection techniques for discrete class data mining, *IEEE Trans. Knowl. Data Eng.* **15** (2003) 1437–1447; DOI: 10.1109/TKDE.2003.1245283.
11. T. Wajima, K. Fukumura, Y. Yano and T. Oguma, Prediction of human clearance from animal data and molecular structural parameters using multivariate regression analysis, *J. Pharm. Sci.* **91** (2002) 2489–2499; DOI: 10.1002/jps.10242.
12. P. P. Roy, S. Paul, I. Mitra and K. Roy, On two novel parameters for validation of predictive QSAR models, *Molecules* **14** (2009) 1660–1701; DOI: 10.3390/molecules 15010604.
13. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim 1999, pp. 125–154.
14. H. Li, Y. Liang and Q. Xu, Support vector machines and its applications in chemistry, *Chemometr. Intell. Lab.* **95** (2009) 188–198; DOI: 10.1016/j.chemolab.2008.10.007.
15. S. K. Shevade, S. S. Keerthi, C. Bhattacharyya and K. R. K. Murthy, *Improvements to SMO Algorithm for SVM Regression,* Technical Report CD-99-16, Control Division, Department of Mechanical and Production Engineering, National University of Singapore, Singapore 1999.
16. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer and V. Steinhauer, Chemical information in 3D space, *J. Chem. Inf. Comput. Sci.* **36** (1996) 1030–1037; DOI: 10.1021/ci960343.
17. M. Petitjean, Applications of the radius-diamater diagram to the classification of topological and geometrical shapes of chemical compounds, *J. Chem. Inf. Comput. Sci.* **32** (1992) 331–337; DOI: 10.1021/ci00008a012.
18. R. S. Obach, J. G. Baxter, T. E. Liston, B. M. Silber, B. C. Jones, F. MacIntyre, D. J. Rance and P. Wastall, The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data, *J. Pharmacol Exp. Ther.* **283** (1997) 46–58.

B. Louis and V. K. Agrawal: Quantitative structure-pharmacokinetic relationship (QSPkR) analysis of the volume of distribution values of anti-infective agents from J group of the ATC classification in humans, *Acta Pharm.* **62** (2012) 305–323.

S A Ž E T A K

Kvantitativni odnos strukture i farmakokinetičkih parametara (QSPkR) volumena distribucije antiinfektivnih lijekova

BRUNO LOUIS i VIJAY K. AGRAWAL

U radu je određen kvantitativni odnos strukture i farmakokinetičkih parametara (QSPkR) za volumen distribucije (V_d) 126 antiinfektivnih lijekova u ljudi koristeći višestruku linearnu regresiju (MLR), umjetne neuronske mreže (ANN), regresiju potpornim vektorima (SVM) i teorijske molekulske deskriptore. Selekcija na temelju korelacije (CFS) upotrjebljena je za izbor relevantnih deskriptora za modeliranje. Rezultati su pokazali da su glavni faktori koji utječu na V_d antiinfektivnih lijekova 3D molekulski prikaz van der Waalsovih volumena atoma i Sandersonove elektronegativnosti, broj alifatskih i aromatskih skupina, broj beta-laktamskih prstena i topološki 2D oblik molekule. Prediktivnost modela procijenjena je vanjskom validacijom, koristeći različite statističke testove. SVM model pokazao se boljim od ostalih modela. Razvijeni model može se upotrijebiti za predviđanje vrijednosti V_d antiinfektivnih lijekova.

Ključne riječi: QSPkR, QSPR, odnos strukture i farmakokinetičkih parametara, volumen distribucije, ANN, SVM, CFS

Department of Pharmacy, Sultan Qaboos University Hospital, PO Box 38, Al Khod, Muscat 123, Oman

QSAR and Computer Chemical Laboratories, A.P.S. University, Rewa-486003, India