Research article

Drug Testing and Analysis

Received: 12 December 2010 Revised: 28 December 2010 Accepted: 30 December 2010 Published online in Wiley Online Library: 25 March 2011

(www.drugtestinganalysis.com) DOI 10.1002/dta.262

Determination of partitioning of drug molecules using immobilized liposome chromatography and chemometrics methods

Hadi Noorizadeh* and Abbas Farmany

The quantitative structure-property relationship (QSPR) of drug molecules against the immobilized liposome chromatography partitioning (log K_s) was studied. The genetic algorithm (GA) was employed to select the variables that resulted in the best-fitted models. After the variables were selected, the linear multivariate regressions (e.g. partial least squares (PLS)) as well as the non-linear regressions (e.g. the kernel PLS (KPLS) and Levenberg-Marquardt artificial neural network (L-M ANN)) were utilized to construct the linear and non-linear QSPR models. The correlation coefficient cross validation (Q^2) and relative error for calibration, prediction and test sets L-M ANN model are (0.987, 0.971, 0.952) and (3.14, 3.54, 6.61), respectively. The obtained results using L-M ANN were compared with those of GA-PLS and GA-KPLS, exhibiting that the L-M ANN model demonstrated a better performance than that of the other models. This is the first research on the QSPR of the drug molecules against the log K_s using the GA-KPLS and L-M ANN.

Keywords: drug molecules; immobilized liposome chromatography; QSPR; genetic algorithm-kernel partial least squares; Levenberg-Marquardt artificial neural network.

Introduction

Because liposomes possess a lipid bilayer structure which mimics biological membranes, they have been used as model membranes for the study of interactions between membranes and biologically important molecules, such as proteins, peptides, and drugs. To quantify the interactions between drugs or other small solutes and liposomal membranes, equilibrium membrane partition coefficients have frequently been determined by a sedimentation method using large multilamellar liposomes.[1] Immobilized liposome chromatography (ILC) has recently been developed as a convenient and rapid method for analysis of solute–membrane interactions.[2] An important application of ILC is the prediction of drug-membrane transport or the analysis of solute–membrane interactions.[3–6]

The drug partitioning was assessed from the retention volume, which was expressed as a capacity factor, K_s, normalized with respect to the amount of immobilized phospholipids, and can be used for semi-quantitative analysis of solute–membrane interaction.[3,4] For normalization of the results obtained on gel beds with different amounts of phospholipids, and for elimination of the dead volume of the system, a logarithm of the normalized retention values, K_s, for a drug was calculated by using Eqn 1:

$$K_s = (V_r - V_o)/A \quad (1)$$

where V_r is the retention volume of the drug, V_o is the retention volume of a small and hydrophilic reference ion (whereby the dead volume of the system is eliminated) and A is the amount of immobilized phospholipids.[7] Using chemometric tools to predict drug and chemical tissue distribution, membrane permeability or biphasic system partition is of major importance in physicochemical, environmental, and life sciences.[8] Quantitative structure–property relationship (QSPR) techniques based on different molecular descriptors have been successfully used to model organic chemicals partition properties.[9]

Computational drug design is a rapidly growing field and an important component of the medicinal chemistry discipline. It is aimed at shortening the drug discovery process which otherwise may be long and expensive. Computationally determined retention parameters have become crucial in identifying potential drug candidates, and this technique is used in lead and clinical candidate optimization, as well as in the selection of new compounds for screening. A number of reports dealing with QSPR calculation of several compounds have been published in the literature.[10–12]

The QSPR models apply to partial least squares (PLS) methods often combined with genetic algorithms (GA) for feature selection.[13,14] Because of the complexity of relationships between the property of molecules and structures, non-linear models are also used to model the structure–property relationships. Levenberg-Marquardt artificial neural network (L-M ANN) is a non-parametric, non-linear modelling technique that has attracted increasing interest. In recent years, non-linear, kernel-based algorithms such as kernel partial least squares (KPLS) have been proposed.[15–17] KPLS can efficiently compute latent variables in the feature space by means of integral operators and non-linear kernel functions. Compared to other non-linear methods, the main advantage of the kernel-based algorithm is that it does not involve non-linear optimization. In the present study, GA-PLS, GA-KPLS, and L-M ANN were employed to generate QSPR models that correlate the structure of some drugs, with observed partitioning on liposome columns (log K_s). The present study is a first research on

* *Correspondence to: Hadi Noorizadeh, Faculty of Sciences, Islamic Azad University, Ilam Branch, Ilam, Iran. E-mail: hadinoorizadeh@yahoo.com*

Faculty of Sciences, Islamic Azad University, Ilam Branch, Ilam, Iran

QSPR of the drug molecules against the log K_s, using GA-KPLS and L-M ANN.

Experimental

Data set

In the current research, the data set was taken from the Osterberg *et al*. [18] The partitioning of a chemically diverse set of drugs into liposomes was studied by ILC. The partitioning was calculated as a capacity factor from the retention volume and provides information about the interaction between the lipids and the substances studied. The drug partitioning was normalized with respect to the amount of immobilized phospholipid. The liposomes composed of synthetic phosphatidylcholine–synthetic phosphatidylethanolamine (PC-PE) 80 : 20 (mol/mol) were immobilized in gel beads by freeze–thawing. The drugs comprised homologous series of compounds such as β-adrenoceptor blockers, local anaesthetics and steroids as well as a set of chemically diverse compounds. A complete list of the drugs' names and their corresponding experimental log K_s is given in Table 1. The partitioning of these compounds was decreased in the range of 3.38 and 0.27 for both Tolfenamic acid and Cephalexin, respectively.

Descriptor calculation

All structures of compounds were drawn with the HyperChem 6.0 program. The pre-optimization of all molecules was performed using MM+ molecular mechanics force field. A more precise optimization was done using the semi-empirical AM1 method in HyperChem. The molecular structures were optimized using the Fletcher-Reeves algorithm until the root mean square gradient was 0.01. Since the calculated values of the quantum chemical features of molecules will be influenced by the related conformation, in the current research an attempt was made to use the most stable conformations. Some quantum chemical descriptors, such as polarizibility and orbital energies of LUMO (lowest unoccupied molecular orbital) and HOMO (highest occupied molecular orbital) were calculated by using the HyperChem program. The output files were transferred into the DRAGON 3.0 program to calculate 1497 molecular descriptors.[19]

Genetic algorithm

A detailed description of the GA can be found in the literature.[20–22] GA is a simulated method based on ideas from Darwin's theory of natural selection and evolution (the struggle for life). In GA, a chromosome (or an individual) can be defined as an enciphered entity of a candidate solution, which is expressed as a set of variables. A GA consists of the following basic steps: (1) a chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) a value for the fitness function of each chromosome is evaluated; and (3) based on the values of the fitness functions, the chromosomes of the next generation are produced by selection, crossover, and mutation operations. The fitness function was proposed by Depczynski *et al*. [23] The parameter algorithm is reported in Table 2.

Table 1. The data set and the corresponding observed and predicted log K_s values by L-M ANN for the calibration, prediction, and test sets

No.	Name	log $K_{s\ Exp}$	log $K_{s\ L-M\ ANN}$	RE(%)
	Calibration set			
1	Cephalexin	0.27	0.25	7.41
2	Atenolol	0.65	0.63	3.08
3	Practolol	0.70	0.68	2.86
4	Inogatran	0.74	0.73	1.35
5	Nadolol	0.86	0.91	5.81
6	Sulpiride	0.89	0.88	1.12
7	5-Phenylvaleric	0.95	0.95	0.00
8	Acebutolol	1.00	1.04	4.00
9	Lidocaine	1.07	1.11	3.74
10	Metoprolol	1.10	1.15	4.55
11	Salicylic acid	1.19	1.17	1.68
12	Ketoprofen	1.25	1.26	0.80
13	Indoprofen	1.30	1.28	1.54
14	Tolmetin	1.36	1.33	2.21
15	Oxprenolol	1.54	1.49	3.25
16	Bupivacaine	1.55	1.54	0.65
17	Sulindac	1.59	1.63	2.52
18	Pindolol	1.69	1.69	0.00
19	4-Phenylbutyl	1.76	1.73	1.70
20	Piroxicam	1.87	1.75	6.42
21	Tetrapeptide	1.87	1.83	2.14
22	Flurbiprofen	2.03	1.96	3.45
23	Furosemide	2.08	2.07	0.48
24	5-Hydroxyquinoline	2.09	2.13	1.91
25	Tetracaine	2.26	2.38	5.31
26	Alprenolol	2.29	2.26	1.31
27	Dexamethasone	2.38	2.28	4.20
28	Testosterone	2.44	2.51	2.87
29	Verapamil	2.56	2.78	8.59
30	Diazepam	2.59	2.72	5.02
31	Indomethacin	2.63	2.68	1.90
32	Propranolol	2.73	2.80	2.56
33	Oxazepam	2.80	2.61	6.79
34	Mefenamic acid	2.87	2.72	5.23
35	Desmethyldiazepam	2.94	2.86	2.72
36	Diflunisal	2.96	2.91	1.69
37	Loperamide	3.28	3.16	3.66
38	Promethazine	3.33	3.48	4.50
39	Tolfenamic acid	3.38	3.50	3.55
	Prediction set			
40	Amoxicillin	0.47	0.49	4.26
41	Theophylline	0.82	0.85	3.66
42	Procaine	1.08	1.06	1.85
43	Terbutaline	1.20	1.22	1.67
44	Naproxen	1.37	1.46	6.57
45	Ibuprofen	1.56	1.62	3.85
46	Fenbufen	1.97	2.06	4.57
47	Tetrapeptide	2.07	2.03	1.93
48	Corticosterone	2.14	2.21	3.27
49	Metolazone	2.32	2.31	0.43
50	Olsalazine	2.65	2.76	4.15
51	Diclofenac	2.77	2.86	3.25
52	Flufenamic acid	3.22	3.01	6.52
	Test set			
53	Amlodipine	0.65	0.68	4.62
54	AVP	0.97	0.91	6.19

Table 1. (*Continued*)

No.	Name	log $K_{s\ Exp}$	log $K_{s\ L-M\ ANN}$	RE(%)
55	d-DAVP	1.08	1.20	11.11
56	Prilocaine	1.19	1.12	5.88
57	Warfarin	1.69	1.82	7.69
58	Omeprazole	1.87	1.96	4.81
59	Hydrocortisone	1.97	2.15	9.14
60	Gemfibrozil	2.03	2.00	1.48
61	Phenytoin	2.59	2.42	6.56
62	Sulphasalazine	2.77	3.01	8.66

Table 2. Parameters of the genetic algorithm

Population size: 30 chromosomes
On average, five variables per chromosome in the original population
Regression method: PLS, KPLS
Cross validation: leave-group-out
Number subset: 4
Maximum number of variables selected in the same chromosome: (PLS, 30)
Elitism: True
Crossover: multi Point
Probability of crossover: 50%
Mutation: multi Point
Probability of mutation: 1%
Maximum number of components: (PLS, 10)
Number of runs: 100

Non-linear models

Kernel partial least squares

The KPLS method is based on the mapping of the original input data into a high-dimensional feature space $\Im$ where a linear PLS model is created. By non-linear mapping $\Phi : x \in \Re^n \rightarrow \Phi(x) \in \Im$, a KPLS algorithm can be derived from a sequence of NIPALS (non-linear iterative partial least squares) steps and has the following formulation:[24–26]

1. Initialize score vector as equal to any column of Y.
2. Calculate scores $u = \Phi\Phi^T w$ and normalize u to $||u|| = 1$, where Φ is a matrix of regressors.
3. Regress columns of Y on u: $c = Y^T u$, where c is a weight vector.
4. Calculate a new score vector w for Y: w = Yc and then normalize w to $||w|| = 1$.
5. Repeat steps 2–4 until convergence of w.
6. Deflate $\Phi\Phi^T$ and Y matrices:

$$\Phi\Phi^T = (\Phi - uu^T\Phi)(\Phi - uu^T\Phi)^T \quad (2)$$

$$Y = Y - uu^T Y \quad (3)$$

7. Go to step 1 to calculate the next latent variable.

Without explicitly mapping into the high-dimensional feature space, a kernel function can be used to compute the dot products as follows:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \quad (4)$$

$\Phi\Phi^T$ represents the (n×n) kernel Gram matrix K of the cross dot products between all mapped input data points $\Phi(x_i), j = 1, \ldots, n$.

The deflation of the $\Phi\Phi^T = K$ matrix after extraction of the u components is given by:

$$K = (I - uu^T)K(I - uu^T) \quad (5)$$

where I is an m-dimensional identity matrix. Taking into account the normalized scores, u, of the prediction of KPLS, the model on training data $\hat{Y}$ is defined as:

$$\hat{Y} = KW(U^T KW)^{-1} U^T Y = -UU^T Y \quad (6)$$

For predictions on new observation data $\hat{Y}_t$, the regression can be written as:

$$\hat{Y}_t = K_t W(U^T KW)^{-1} U^T Y \quad (7)$$

where K_t is the test matrix whose elements are $K_{ij} = K(x_i, x_j)$ where x_i and x_j present the test and training data points, respectively.

Artificial neural network

An artificial neural network (ANN) with a layered structure is a mathematical system that stimulates the biological neural network; it consists of computing units named neurons and connections between neurons named synapses.[27,29] Input or independent variables are considered as neurons of input layer, while dependent or output variables are considered as output neurons. Synapses connect input neurons to hidden neurons and hidden neurons to output neurons. The strength of the synapse from neuron *i* to neuron *j* is determined by mean of a weight, W_{ij}. In addition, each neuron *j* from the hidden layer, and eventually the output neuron, is associated with a real value b_j, named the neuron's bias and with a non-linear function, named the transfer or activation function. Because ANNs are not restricted to linear correlations, they can be used for non-linear phenomena or curved manifolds.[27] Back propagation neural networks (BNNs) are most often used in analytical applications.[28] The back propagation network receives a set of inputs, which is multiplied by each node and then a non-linear transfer function is applied. The goal of training the network is to change the weight between the layers in a direction to minimize the output errors. The changes in values of weights can be obtained using Eqn (10):

$$\Delta W_{ij,n} = F_n + \alpha \Delta W_{ij,n-1} \quad (8)$$

whereis ΔW_{ij} the change in the weight factor for each network node, α is the momentum factor, and *F* is a weight update function, which indicates how weights are changed during the learning process. There is no single best weight update function which can be applied to all non-linear optimizations. One needs to choose a weight update function based on the characteristics of the problem and the data set of interest. Various types of algorithms have been found to be effective for most practical purposes such as the Levenberg-Marquardt (L-M) algorithm.

Levenberg-Marquardt algorithm. While basic back propagation is the steepest descent algorithm, the Levenberg-Marquardt algorithm[30] is an alternative to the conjugate methods for second derivative optimization. In this algorithm, the update function, F_n, can be calculated using Eqns (11) and (12):

$$F_0 = -g_0 \quad (9)$$

$$F_n = -[J^T \times J + \mu I]^{-1} \times JT \times e \quad (10)$$

where J is the Jacobian matrix, μ is a constant, *I* is an identity matrix, and *e* is an error function.[31]

Software and programs

A Pentium IV personal computer (CPU at 3.06 GHz) with Windows XP operational system was used. Geometry optimization was performed by HyperChem (Version 7.0 Hypercube, Inc.); Dragon software was used to calculate of descriptors. MINITAB software (version 14, MINITAB) was used for the simple PLS analysis. Cross validation, GA-PLS, GA-KPLS, L-M ANN and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

Results and discussion

Linear model

Results of the GA-PLS model

To reduce the original pool of descriptors to an appropriate size, the objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. After this process, 1003 descriptors remained. These descriptors were employed to generate the models with the GA-PLS and GA-KPLS program. The best model was selected on the basis of the highest multiple correlation coefficient leave-group-out cross validation (LGO-CV) (Q^2), the least root mean squares error (RMSE) and relative error (RE) of prediction and simplicity of the model. These parameters are probably the most popular measures of how well a model fits the data.

Among the models proposed by GA-PLS, one model had the highest statistical quality and was repeated more than the others. The best GA-PLS model contains 7 selected descriptors in 3 latent variables space. These descriptors were obtained constitutional descriptors ((sum of atomic polarizabilities (scaled on carbon atom (SP)) and sum of atomic van der Waals volumes (scaled on carbon atom) (Sv)), geometrical descriptors (SPAN (span R)), 3D-MoRSE descriptors (3D-MoRSE - signal 07/weighted by atomic Sanderson electronegativities (Mor07e)), molecular properties (Squared Moriguchi octanol-water partition coeff. (logP^2) (MLOGP2)) and quantum chemical descriptors (polarizibility and lowest unoccupied molecular orbital (LUMO)). For this in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis. The Q^2, mean relative error and RMSE for training and test sets were (0.881, 0.752), (9.6, 18.16) and (7.90, 16.25), respectively. The PLS model uses a higher number of descriptors that allows the model to extract better structural information from descriptors to result in a lower prediction error.

Non-linear models

Results of the GA-KPLS model

The LGO-CV was performed. In this paper, a radial basis kernel function, $k(x,y) = \exp(||x-y||^2/c)$, was selected as the kernel function with $c = rm\sigma^2$ where r is a constant that can be determined by considering the process to be predicted (here r was set to be 1), m is the dimension of the input space and σ^2 is the variance of the data.[32] It means that the value of c depends on the system under the study. The 6 descriptors in 3 latent variables space chosen by GA-KPLS feature selection methods were contained. These descriptors were obtained constitutional descriptors (number of Hydrogen atoms (nH)), charge descriptors (relative negative charge (RNCG) and total positive charge (Qpos)), molecular properties (topological polar surface area using N, O, S, P polar contributions (TPSA (Tot)) and Moriguchi octanol-water partition coefficients (logP) (MLOGP)) and quantum chemical descriptors (HOMO)). The Q^2, mean RE, and RMSE for training and test sets were (0.929, 0.899), (5.78, 9.86) and (3.04, 6.70), respectively. The RMSE values of the GA-KPLS model for the training and test sets were much lower than the GA-PLS model. From these results, it can be noticed that the GA-KPLS model gave the highest Q^2 values, so this model provided the most satisfactory results, compared with the results obtained from the GA-PLS model. The GA-PLS linear model had good statistical quality with low prediction error, while the corresponding errors obtained by the GA-KPLS model were lower. Consequently, this GA-KPLS approach currently constitutes the most accurate method for predicting the partitioning of the drug components rather than the GA-PLS method. This suggests that GA-KPLS hold promise for applications in choosing a variable for L-M ANN systems. This result indicates that the log K_s of drug molecules possesses some non-linear characteristics.

Results of the L-M ANN model

With the aim of improving the predictive performance of non-linear QSPR model, L-M ANN modelling was performed. Descriptors of GA-KPLS model were selected as inputs in L-M ANN model. The network architecture consisted of 6 neurons in the input layer corresponding to the 6 mentioned descriptors. The output layer had 1 neuron that predicted the log K_s. The number of neurons in the hidden layer was unknown and needed to be optimized. In addition to the number of neurons in the hidden layer, the learning rate, the momentum and the number of iterations should also be optimized. In this work, the number of neurons in the hidden layer and other parameters except the number of iterations were simultaneously optimized. A MATLAB program was written to change the number of neurons in the hidden layer from 2 to 7, the learning rate from 0.001 to 0.1 with a step of 0.001, and the momentum from 0.1 to 0.99 with a step of 0.01. The RMSE for training set were calculated for all possible combination of values for the mentioned variables in LGO-CV. It was realized that the RMSE for the training set was minimum when two neurons were selected in the hidden layer and the learning rate and the momentum values were 0.5 and 0.3, respectively. Finally, the number of iterations was optimized with the optimum values for the variables. It was realized that after 14 iterations, the RMSE for prediction set was minimum. The values of experimental, calculated, and percent relative error are shown in Table 1. The Q^2, mean RE, and RMSE for calibration, prediction and test sets were (0.986, 0.971, 0.952), (3.11, 3.54, 6.61), and (0.07, 0.08, 0.13), respectively. For the constructed model, three general statistical parameters were selected to evaluate the prediction ability of the model for the log K_s. The statistical parameters Q^2, RE, and RMSE were obtained for the proposed models. Each of the statistical parameters mentioned above was used for assessing the statistical significance of the QSPR model. Inspection of the results revealed a higher Q^2 and lowered other value parameters for the training and test sets compared with their counterparts for GA-KPLS and GA-PLS. Plots of predicted log K_s versus experimental log K_s values by L-M ANN for training and test sets are shown Figures 1A and 1B. Obviously, there is a close agreement between the experimental and predicted log K_s and the data represent a very low scattering around a straight line with respective slope and intercept close

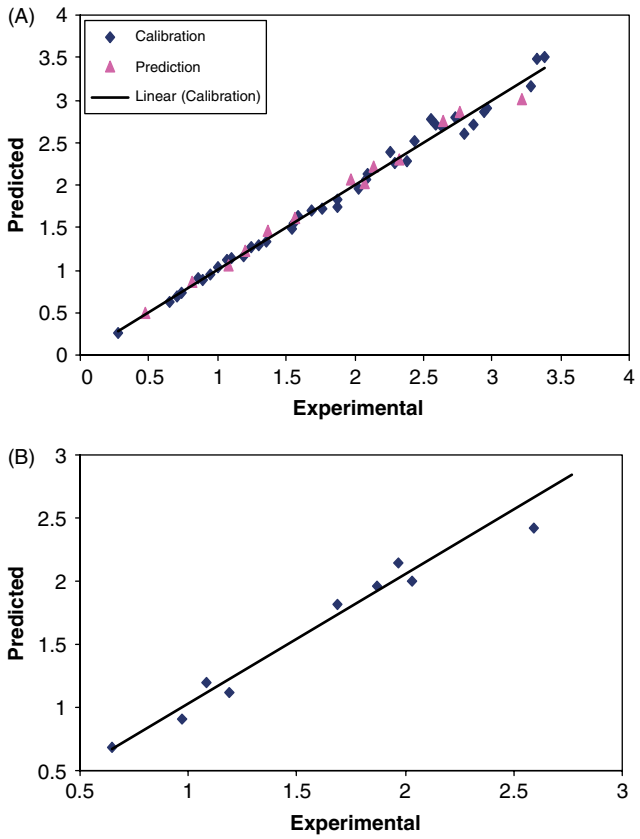

Figure 1. Plot of predicted log K_s obtained by L-M ANN against the experimental values (A) training set of molecules and (B) for test set.

to one and zero. This clearly shows the strength of L-M ANN as a non-linear feature selection method. The key strength of L-M ANN is their ability to allow for flexible mapping of the selected features by manipulating their functional dependence implicitly. Neural network handles both linear and non-linear relationships without adding complexity to the model. This capacity offset the large computing time required and complexity of L-M ANN model with respect to other models. It is easy to notice that there was a good prospect for the L-M ANN application in the QSPR modelling. Three methods seemed to be useful, although a comparison between methods revealed the slight superiority of the L-M ANN to other models. High correlation coefficients training and test set and low prediction errors confirmed the good predictability of the three models. The L-M ANN model can be effectively used to describe the molecular structure characteristic of these drugs. It can also be used successfully to estimate the log K_s for new drug compounds or for other drugs whose experimental values are unknown compared to other models. The advantages of this work are that the number of the used descriptors was smaller and that all the used descriptors were known to be important in the log K_s of drug compounds and also this is the first study on the QSPR of the drug compounds using the GA-KPLS and L-M ANN models.

Model validation

Validation is a crucial aspect of any QSPR/QSRR modelling.[33] The accuracy of the proposed models was illustrated using the evaluation techniques such as LGO-CV procedure and validation through an external test set.

Cross validation technique

Cross validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets were created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model was developed, based on the utilized modelling technique. Each model was evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones or group data that were not utilized in the development of the model).[34] In particular, the LGO-CV procedure was utilized in this study. A QSPR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data. This procedure was repeated until a complete set of predicted was obtained. The statistical significance of the screened model was judged by the correlation coefficient (Q^2). The predictive ability was evaluated by the cross validation coefficient (Q^2 or R^2_{cv}) which is based on the prediction error sum of squares (PRESS) and was calculated by the following equation:

$$R^2_{cv} \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (11)$$

where y_i, $\hat{y_i}$ and $\bar{y}$ were respectively the experimental, predicted, and mean log K_s values of the samples. The accuracy of cross validation results is extensively accepted in the literature considering the Q^2 value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) was considered as proof of the high predictive ability of the model.[35] However, this assumption is, in many cases, incorrect and the lack of the correlation between the high Q^2 and the high predictive ability of QSPR/QSRR models has been established and corroborated recently.[33] Thus, the high value of Q^2 appears to be necessary but not a sufficient condition for the models to have a high predictive power. These authors stated that an external set is necessary. As a next step, further analysis was also followed for chemical property of the new set of compounds using the developed QSPR model.

Validation through the external test set

Validating QSPR with external data (i.e. data not used in the model development) is the best method of validation. However, the availability of an independent external test set of several compounds is rare in QSPR. Thus, the predictive ability of a QSPR model with the selected descriptors was further explored by dividing the full data set. The predictive power of the models developed on the selected training set was estimated on the predicted values of test set chemicals. The data set was randomly divided into three groups including calibration and prediction sets (training set) and test set, which consisted of 39, 13, and 10 molecules, respectively. The calibration set was used for model generation. The prediction set was applied with overfitting of the network, whereas the test set, whose molecules had no role in model building, was used for the evaluation of the predictive ability of the models for external set. The result clearly displays a significant improvement of the QSPR model consequent to non-linear statistical treatment and a substantial independence of model prediction from the structure of the test molecule. In the above analysis, the descriptive power of a given model was

measured by its ability to predict partition of unknown drugs. For instance, as to prediction ability, it can be observed in Figure 1 that scattering of data points from the ideal trend in test set is poor.

Interpretation of descriptors

Liposomes possess an orderly molecular structure and are able to exert electrostatic interactions.[36,37] Generally, the membrane partitioning of drugs decreased in the order neutral ⟩ positively charged ⟩ negatively charged drugs. Possibly the partitioning of neutral drugs into the bilayers is stronger than that of charged drugs, whereas the combined electrostatic and hydrophobic effect between positively charged drugs and the bilayer is stronger than that between negatively charged drugs and the bilayer, since the positively charged drugs spend more time imbedded in the hydrocarbon region owing to interaction with the adjacent phosphate group, whereas the negatively charged drugs tend to be dragged out of the bilayer due to electrostatic attraction to the positive charges in the outermost parts of the head groups. The inclusion of a negatively charged PC-PE liposome increased the retention of positively charged drugs and decreased the retention of negatively charged drugs.[18]

The partitioning of drugs on liposome columns depends upon a number of molecular properties, such as lipophilicity, molecular size, polarity, charge and molar volume.[18] The effect of the charge on the retention in the ILC columns in part may be explained by different molar volumes of the compounds.

Because a cell membrane is comprised of hydrophilic and lipophilic regions, a molecule that passes through a cell membrane through the transcellular pathway needs to penetrate both hydrophilic and hydrophobic environments. As a result, both hydrophilic and lipophilic properties of a drug should be taken into account when predicting drug permeability. It is difficult for a drug molecule with a mainly hydrophilic structure to penetrate the outer layer (phospholipids layer) of the cell membrane by transcellular diffusion.

Log *P* is a quantitative descriptor of lipophilicity and estimates the propensity of a neutral compound to differentially dissolve in two immiscible phases. Lipophilicity is approximately correlated to passive transport across cell membranes and the ability of a compound to partition through a membrane since membranes are composed largely of lipids. It is usually referred to the octanol–water partition coefficient (*P*), expressed as a logarithmic ratio. Nowadays, log *P* is commonly used in QSPR/QSAR study and drug design since it relates to drug absorption, bioavailability, metabolism, and toxicity.[38]

Constitutional descriptors are the simplest and most commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry. The most common constitutional descriptors are number of atoms, number of bound, absolute and relative numbers of specific atom type, absolute and relative numbers of single, double, triple, and aromatic bound, number of ring, number of ring divided by the number of atoms or bonds, number of benzene ring, number of benzene ring divided by the number of atoms, molecular weight and average molecular weight.

Hydrogen bonding is a measure of the tendency of a molecule to form hydrogen bonds. This is related to number of hydrogen atoms (nH). Hydrogen bonding may be divided into an electrostatic term and a polarization/charge transfer term. A particularly strong type of polar interaction occurs in molecules where a hydrogen atom is attached to an extremely electron-hungry atom such as oxygen, nitrogen, or fluorine. Understandably, hydrogen bonding plays a significant role in retention behaviour.

3D-MoRSE (3D-MOlecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies. These descriptors are calculated by summing atom weights viewed by a divergent angular scattering function.

Geometrical descriptors encode information about the 3D features (size and shape) of a molecule. They depend on the three-dimensional coordinates determined by molecular modelling. They are usually employed only in situations in which locked conformations are being studied. These descriptors attempt to describe the geometrical environments of carbon atoms. The SPAN is a size descriptor definite as the radius of the smallest sphere, centred on the centre of mass, completely enclosing all atoms of a molecule.

Although lipophilicity, molecular volume, geometrical descriptors, and molar volume are often successful in rationalizing partition of drugs on liposome columns, they cannot account for conformational changes and they do not provide information about electronic influence through bonds or across space. For that reason, quantum chemical descriptors are used in developing QSPR.

Quantum chemical descriptors can give great insight into structure and reactivity and can be used to establish and compare the conformational stability, chemical reactivity, and intermolecular interactions. They include thermodynamic properties (system energies) and electronic properties (LUMO or HOMO energy). Quantum chemical descriptors were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such as atoms, bonds, and molecular fragments. Electronic properties may play a role in the magnitude of a biological activity, along with structural features encoded in indexes. The eigen values of LUMO and HOMO and their energy gap reflect the chemical activity of the molecule. LUMO as an electron acceptor represents the ability to obtain an electron, while HOMO as an electron donor represents the ability to donate an electron. The HOMO energy plays a very important role in the nucleophylic behaviour and it represents molecular reactivity as a nucleophyle. Good nucleophyles are those where the electron residue is high lying orbital. The energy of the LUMO is directly related to the electron affinity and characterizes the susceptibility of the molecule towards attack by nucleophiles. The LUMO energy can be interpreted as a measure of charge transfer interactions and/or of hydrogen-bonding effects. Electron affinity was also shown to greatly influence the chemical behaviour of compounds, as demonstrated by its inclusion in the QSPR.

Polar functional groups account for many of the dipole–dipole, dipole-induced dipole and hydrogen bond interactions. Drugs with high polarity are less likely to be absorbed from the small intestine. Topological polar surface area (TPSA) also accounts for the steric shape of a molecule and has been found to be related to drug permeability. The TPSA is a surface descriptor, defined as the part of the surface area of a molecule contributed by nitrogen, oxygen, and connected hydrogen atoms. As such, it is clearly related to the capacity of a drug to form hydrogen bonds. Molecules with many H-bond donors and a large polar surface area yields low permeability values. It can be observed that, for molecules with large TPSA, permeability increases with lipophilicity, while for molecules with small TPSA, lipophilicity appears to have little effect on intestinal permeability.

Charge descriptors are defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such as atoms, bonds, and molecular fragments. Electrical charges in the molecule are the driving force of electrostatic interactions, and it is well known that local electron densities or charge play a fundamental role in many physic-chemical properties and receptors-ligand binding affinity. Thus, charge-based descriptors have been widely employed as chemical reactivity indices or as measures of weak intermolecular interactions. Many quantum chemical descriptors are derived from the partial charge distribution in a molecule or from the electron densities on particular atoms. Relative negative charge (RNCG) is partial charge of the most negative atom divided by the total negative:[39]

$$RNCG = \frac{Q^{-}_{max}}{Q^{-}} \quad (12)$$

The RNCG descriptor increased the retention of positively charged drugs and decreased the retention of negatively charged drugs.

From the above discussion, it can be seen that the molecular size, hydrogen bond interactions, and electrostatic interactions are the three likely factors controlling the partitioning of these drugs. All the descriptors involved in the model, which have explicit physical meaning, may account for the structure responsible for the partitioning of these compounds.

Conclusion

In the present study, one linear method (GA-PLS) and two non-linear methods (GA-KPLS and L-M ANN) were used to construct a quantitative relation between the partitioning of drugs on liposome columns and their calculated descriptors. The most important molecular descriptors selected represent the molecular properties, constitutional, charge, and quantum chemical descriptors that are known to be important in the retention mechanism of drug molecules. The results obtained by L-M ANN were compared with the results obtained by other models. The results demonstrated that L-M ANN was more powerful in the partitioning prediction of the drug molecules than GA-PLS and GA-KPLS. A suitable model with high statistical quality and low prediction errors was eventually derived. This model could accurately predict the partitioning of these components that did not exist in the modelling procedure. It was easy to notice that there was a good prospect for the L-M ANN application in the QSPR modelling.

References

[1] Y. W. Choi, J. A. Rogers, *Pharmaceut. Res.* **1990**, *7*, 508.
[2] H. Ishii, T. Shimanouchi, H. Umakoshi, R. Kuboi, *J. Biosci. Bioeng.* **2009**, *108*, 425.
[3] N. Yoshimoto, M. Yoshimoto, K. Yasuhara, T. Shimanouchi, H. Umakoshi, R. Kuboi, *Biochem. Eng. J.* **2006**, *29*, 174.
[4] L. H. Sheng, S. L. Li, L. Kong, X. G. Chen, X. Q. Mao, X. Y. Su, H. F. Zou, P. Li, *J. Pharmaceut. Biomed.* **2005**, *38*, 216.
[5] Y. Wang, L. Kong, X. Lei, L. Hu, H. Zou, E. Beck, S. W. A. Bligh, Zh. Wang, *J. Chromatogr. A* **2009**, *1216*, 2185.
[6] C. Huang, J. T. Mason, *P. Natl. Acad. Sci. USA* **1978**, *75*, 308.
[7] E. Boija, A. Lundquist, J. J. Martínez Pla, C. Engvall, P. Lundahl, *J. Chromatogr. A* **2004**, *1030*, 273.
[8] G. Klopman, H. Zhu, *Mini Rev. Med. Chem.* **2005**, *5*, 127.
[9] G. Schuurmann, R. U. Ebert, R. Kuhne, *Environ. Sci. Technol.* **2006**, *40*, 7005.
[10] B. Xia, W. Ma, X. Zhang, B. Fan, *Anal. Chim. Acta* **2007**, *598*, 12.
[11] J. M. Bermudez-Saldana, L. Escuder-Gilabert, M. J. Medina-Hernandez, R. M. Villanueva-Camanas, S. Sagrado, *Chemosphere* **2007**, *69*, 108.
[12] W. Ma, F. Luan, H. Zhang, X. Zhang, M. Liu, Zh. Hu, B. Fan, *J. Chromatogr. A* **2006**, *1113*, 140.
[13] H. Noorizadeh, A. Farmany, *Chromatographia* **2010**, *72*, 563.
[14] H. Golmohammadi, M. Safdari, *Microchem. J.* **2010**, *95*, 140.
[15] S. H. Woo, Ch. O. Jeon, Y. S. Yun, H. Choi, Ch. S. Lee, D. S. Lee, *J. Hazard. Mater.* **2009**, *161*, 538.
[16] N. Krämer, A. L. Boulesteix, G. Tutz, *Chemometr. Intell. Lab.* **2008**, *94*, 60.
[17] S. Haykin, *Neural Networks*, Prentice-Hall: NJ, USA, **1999**.
[18] T. Osterberg, M. Svensson, P. Lundahl, *Eur. J. Pharm. Sci.* **2001**, *12*, 427.
[19] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON-Software for the calculation of molecular descriptors. Version 3.0 for Windows, **2003**.
[20] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning,* Addison-Wesley – Longman: Reading, MA, USA, **2000**.
[21] S. Riahi, E. Pourbasheer, R. Dinarvand, M. R. Ganjali, P. Norouzi, *Chem. Biol. Drug Des.* **2008**, *72*, 205.
[22] H. Noorizadeh, A. Farmany, A. Khosravi, *J. Chin. Chem. Soc.* **2010**, *57*, 1.
[23] U. Depczynski, V. J. Frost, K. Molt, *Anal. Chim. Acta* **2000**, *420*, 217.
[24] S. Wold, M. Sjostrom, L. Eriksson, *Chemometr. Intell. Lab.* **2001**, *58*, 109.
[25] P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1.
[26] R. Rosipal, L. J. Trejo, *J. Mach. Learn. Res.* **2001**, *2*, 97.
[27] J. Zupan, J. Gasteiger, *Neural Network in Chemistry and Drug Design*, Wiley-VCH: Weinheim, Germany, **1999**.
[28] T. M. Beal, H. B. Hagan, M. Demuth, *Neural Network Design*, PWS: Boston, MA, USA, **1996**.
[29] J. Zupan, J. Gasteiger, *Neural Networks for Chemists: An Introduction*, Wiley-VCH: Weinheim, Germany, **1993**.
[30] S. Kara, M. Okandan, *Pattern Recogn.* **2007**, *40*, 2967.
[31] M. Salvi, D. Dazzi, I. Pelistri, F. Neri, J. R. Wall, *Ophthalmology* **2002**, *109*, 1703.
[32] K. Kim, J. M. Lee, I. B. Lee, *Chemometr. Intell. Lab.* **2005**, *79*, 22.
[33] J. Acevedo-Martınez, J. C. Escalona-Arranz, A. Villar-Rojas, F. Tellez-Palmero, R. Perez-Roses, L. Gonzalez, R. Carrasco-Velar, *J. Chromatogr. A* **2006**, *1102*, 238.
[34] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, *Bioorgan. Med. Chem.* **2006**, *14*, 6686.
[35] H. Noorizadeh, A. Farmany, *J. Chin. Chem. Soc.* **2010**, *57*, 1.
[36] M. M. Felix, H. Umakoshi, T. Shimanouchi, M. Yoshimoto, R. Kuboi, *J. Biosci. Bioeng.* **2002**, *93*, 498.
[37] E. Boija, A. Lundquist, J. J. Martínez Pla, C. Engvall, P. Lundahl, *J. Chromatogr. A* **2004**, *1030*, 273.
[38] V. T. Joseph, D. G. Beverly, A. K. Snezana, *Anal. Chim. Acta* **2003**, *485*, 89.
[39] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors,* Wiley-VCH: Weinheim, Germany, **2000**.