Research article

Drug Testing and Analysis

Received: 4 January 2011 Revised: 2 February 2011 Accepted: 2 February 2011 Published online in Wiley Online Library: 25 March 2011

(www.drugtestinganalysis.com) DOI 10.1002/dta.275

Application of genetic algorithm-kernel partial least square as a novel non-linear feature selection method: partitioning of drug molecules

H. Noorizadeh,[a]* S. Sobhan Ardakani,[b] T. Ahmadi,[a] S. S. Mortazavi[c] and M. Noorizadeh[d]

Genetic algorithm (GA) and partial least squares (PLS) and kernel PLS (KPLS) techniques were used to investigate the correlation between immobilized liposome chromatography partitioning (log Ks) and descriptors for 65 drug compounds. The models were validated using leave-group-out cross validation LGO-CV. The results indicate that GA-KPLS can be used as an alternative modelling tool for quantitative structure-property relationship (QSPR) studies.

Keywords: drug molecules; immobilized liposome chromatography; QSPR; genetic algorithm-kernel partial least squares

Introduction

Liposomes possess a lipid bilayer structure which simulates biological membranes, and therefore they are frequently used as models to study interactions between membranes and biological molecules such as proteins, peptides, and drugs. For chromatographic analysis, liposomes and biomembranes have been immobilized in gel beads by steric entrapment,[1] hydrophobic binding,[2] avidin–biotin affinity binding,[3] or covalent attachment.[4] Immobilized liposome chromatography (ILC) developed in recent years is regarded as a powerful tool to study drug–membrane interactions *in vitro*.[5,6] Liposome[7–9] formed by phosphatidylcholine, the main component found in cell membrane, or unilaminar phospholipids,[10–13] were non-covalently or covalently immobilized on soft gel particles or silica particles as chromatographic stationary phase to probe the penetration ability of compounds through biological membranes, which has been considered as one of most important parameters to evaluate their bioactivity.[14–16]

The partitioning is calculated[17,19] as a capacity factor from the retention volume and provides information about the interaction between the lipids and the substances studied. ILC is a convenient method, since the membranes are stable for months and the analysis is quite rapid, which enables extensive substance and parameter screening.[17–19] Retention of the solute was expressed as a specific capacity factor, K_s.[20]

Using chemometrics tools to predict drugs and chemical tissue distribution, membrane permeability, or biphasic system partition is of major importance in physicochemical, environmental, and life sciences. Chemical distribution phenomena depend not only on molecular structure but also on the properties of the system in question.[21] Quantitative structure–property relationship (QSPR) techniques based on different molecular descriptors have been successfully used to model organic chemicals partition properties.[22] A number of reports that deals with QSPR calculation of several compounds have been published in the literature.[23–25]

The QSPR models can be applied to partial least squares (PLS) methods often combined with genetic algorithms (GA) for feature selection.[26,27] 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. In recent years, non-linear kernel-based algorithms as kernel partial least squares (KPLS) have been proposed.[28,29] The basic idea of KPLS is first to map each point in an original data space into a feature space via non-linear mapping and then to develop a linear PLS model in the mapped space. According to Cover's theorem, non-linear data structure in the original space is most likely to be linear after high-dimensional non-linear mapping.[30] Therefore, KPLS can efficiently compute latent variables in the feature space by means of integral operators and non-linear kernel functions. In the present study, GA-PLS and GA-KPLS were employed to generate QSPR models that correlate the structure of some drugs; with observed partitioning on liposome columns (log K_s).

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

a *Islamic Azad University, Ilam Branch, Ilam, Iran*

b *Department of Environment, Islamic Azad University, Hamedan Branch, Hamedan, Iran*

c *Islamic Azad University, Hamedan Branch, Young Researchers Club, Hamedan, Iran*

d *Members of Young Researchers Club, Islamic Azad University, Ilam Branch, Ilam, Iran*

Table 1. The data set and the corresponding observed and predicted log K_s values by GA-KPLS for training and test set

No.	Name	log $K_{s\ Exp}$	log $K_{s\ Cal}$	RE	AbsE
	Training set				
1	Atenolol	0.24	0.22	8.33	0.02
2	Antipyrine	0.31	0.33	6.45	0.02
3	Ciprofloxacin	0.41	0.44	7.32	0.03
4	Sulpiride	0.43	0.41	4.65	0.02
5	Theophylline	0.45	0.50	11.11	0.05
6	AVP	0.48	0.46	4.17	0.02
7	Procaine	0.69	0.67	2.90	0.02
8	d-DAVP	0.73	0.66	9.59	0.07
9	5-Phenylvaleric	0.74	0.75	1.35	0.01
10	Tramadol	0.78	0.81	3.85	0.03
11	Salicylic acid	0.86	0.79	8.14	0.07
12	Terbutaline	0.94	1.00	6.38	0.06
13	Lidocaine	1.01	1.03	1.98	0.02
14	Enalaprilat	1.04	1.01	2.88	0.03
15	Indoprofen	1.05	1.06	0.95	0.01
16	Prilocaine	1.10	1.02	7.27	0.08
17	Sulindac	1.26	1.19	5.56	0.07
18	Tolmetin	1.28	1.40	9.38	0.12
19	Warfarin	1.40	1.36	2.86	0.04
20	Naproxen	1.43	1.48	3.50	0.05
21	Bupivacaine	1.49	1.61	8.05	0.12
22	Ibuprofen	1.59	1.51	5.03	0.08
23	Piroxicam	1.61	1.58	1.86	0.03
24	4-Phenylbutyl	1.73	1.83	5.78	0.10
25	Tetrapeptide	1.74	1.70	2.30	0.04
26	Omeprazole	1.79	1.71	4.47	0.08
27	Hydrocortisone	1.80	1.74	3.33	0.06
28	Furosemide	1.92	2.09	8.85	0.17
29	Fenbufen	2.00	1.90	5.00	0.10
30	Flurbiprofen	2.03	2.08	2.46	0.05
31	Metolazone	2.09	2.08	0.48	0.01
32	Gemfibrozil	2.11	2.27	7.58	0.16
33	Tetracaine	2.17	2.14	1.38	0.03
34	Corticosterone	2.18	2.30	5.50	0.12
35	Alprenolol	2.28	2.22	2.63	0.06
36	Dexamethasone	2.40	2.55	6.25	0.15
37	Testosterone	2.47	2.24	9.31	0.23
38	Olsalazine	2.51	2.46	1.99	0.05
39	Diazepam	2.58	2.74	6.20	0.16
40	Sulphasalazine	2.68	2.66	0.75	0.02
41	Propranolol	2.70	2.79	3.33	0.09
42	Oxazepam	2.79	2.59	7.17	0.20
43	Diclofenac	2.85	2.86	0.35	0.01
44	Mefenamic acid	2.90	2.72	6.21	0.18
45	Desmethyldiazepam	2.94	3.20	8.84	0.26
46	Loperamide	3.17	2.94	7.26	0.23
47	Cyclosporin	3.18	3.02	5.03	0.16
48	Flufenamic acid	3.24	3.31	2.16	0.07
49	Tolfenamic acid	3.39	3.18	6.19	0.21
50	Promethazine	3.36	3.40	1.19	0.04
51	Proxicromil	3.83	3.43	10.44	0.40
52	Amlodipine	4.02	3.71	7.71	0.31
	Test set				
53	Practolol	0.36	0.31	13.89	0.05
54	Nadolol	0.68	0.75	10.29	0.07
55	Acebutolol	0.84	0.81	3.57	0.03
56	Metoprolol	1.02	1.12	9.80	0.10
57	Ketoprofen	1.17	1.03	11.97	0.14
58	Pindolol	1.43	1.39	2.80	0.04
59	5-Hydroxyquinoline	1.74	1.88	8.05	0.14
60	Tetrapeptide	2.07	2.04	1.45	0.03
61	Verapamil	2.46	2.82	14.63	0.36
62	Phenytoin	2.54	2.31	9.06	0.23
63	Indomethacin	2.72	2.42	11.03	0.30
64	Diflunisal	2.99	3.12	4.35	0.13
65	Salmeterol	3.54	3.92	10.73	0.38

Computational

Data set

In the current research, the data set was taken from the reference.[31] The partitioning of a chemically diverse set of drugs into liposomes was studied by ILC. The partitioning is 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 purified egg phospholipids (EPL) were immobilized in gel beads by freeze–thawing. The drugs comprise 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 4.02 and 0.24 for both Amlodipine and Atenolol, 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 with the semi-empirical AM1 method in HyperChem. The molecular structures were optimized using the Fletcher- Reeves algorithm until the root mean square gradientwas 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 orbital energies of the lowest unoccupied molecular orbital (LUMO) and the highest occupied molecular orbital (HOMO) were calculated by using the HyperChem program. The output files were transferred into the DRAGON 3.0 program to calculate 1497 molecular descriptors.[32]

Genetic algorithm

A detailed description of the GA can be found in the literature.[33–35] The GA is a simulated method based on ideas from Darwin's theory of natural selection and evolution (the struggle for life). In the 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. The GA consists of the following basic steps: (1) a chromosome is represented by a binary bit string and an initial

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

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*.[36] The parameter algorithm reported in Table 2.

Linear model

Partial least squares (PLS)

PLS is a linear multivariate method for relating the process variables X with responses Y. PLS can analyze data with strongly collinear, noisy, and numerous variables in both X and Y.[37] PLS reduces the dimension of the predictor variables by extracting factors or latent variables that are correlated with Y while capturing a large amount of the variations in X. This means that PLS maximizes the covariance between matrices X and Y. In PLS, the scaled matrices X and Y are decomposed into score vectors (t and u), loading vectors (p and q), and residual error matrices (E and F):

$$X = \sum_{i=1}^{a} t_i p_i^T + E$$
$$Y = \sum_{i=1}^{a} u_i q_i^T + F \quad (1)$$

where a is the number of latent variables. In an inner relation, the score vector t is linearly regressed against the score vector u:

$$U_i = b_i t_i + h_i \quad (2)$$

where b is regression coefficient that is determined by minimizing the residual h. It is crucial to determine the optimal number of latent variables; cross validation is a practical and reliable way to test the predictive significance of each PLS component. There are several algorithms to calculate the PLS model parameters. In this work, the non-linear iterative partial least squares (NIPALS) algorithm was used with the exchange of scores.[38]

Non-linear model

Kernel partial least squares (KPLS)

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 steps and has the following formulation:[39]

1. Initialize score vector w as equal to any column of Y.
2. Calculate scores $u = \Phi\Phi^T w$ and normalize u to $||u|| = 1$, where 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:

$$\Phi\Phi^T = (\Phi - uu^T\Phi)(\Phi - uu^T\Phi)^T \quad (3)$$
$$Y = Y - uu^T Y \quad (4)$$

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:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \quad (5)$$

$\Phi\Phi^T$ represents the (n×n) kernel Gram matrix K of the cross dot products between all mapped input data points $\Phi(x_i), i = 1, \ldots, n$. The deflation of the $\Phi\Phi^T = K$ matrix after extraction of the u components is given by:

$$K = (I - uu^T)K(I - uu^T) \quad (6)$$

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

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

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

$$\hat{Y}_t = K_t W(U^T KW)^{-1} U^T Y \quad (8)$$

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.

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 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

Table 3. The statistical parameters of different constructed QSPR models

	Training set						Test set					
Model	R^2	Q^2	RE	RMSE	AbsE	N	R^2	Q^2	RE	RMSE	AbsE	N
GA-PLS	0.871	0.873	9.86	0.24	0.22	52	0.796	0.794	18.09	0.37	0.31	13
GA-KPLS	0.976	0.977	5.01	0.12	0.09	52	0.952	0.951	8.58	0.19	0.15	13

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 is 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), absolute error (AbsE), and relative error (RE) of prediction and simplicity of the model. These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains six selected descriptors in three latent variables space. These descriptors were obtained constitutional descriptors (number of atoms (nAT) and mean atomic polarizability (scaled on Carbon atom) (Mp)), WHIM descriptors (1st component size directional WHIM index/weighted by atomic masses (L1m)), charge descriptors (maximum negative charge (qnmax)), molecular properties (Squared Moriguchi octanol-water partition coeff. (logP) (MLOGP)) and quantum chemical descriptors (HOMO). For this in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis. The obtained statistic parameters of the GA-PLS model are shown in Table 3. 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 model

Results of the GA-KPLS model

With the aim of improving the predictive performance of non-linear QSPR model, GA-KPLS modelling was performed. The LGO-CV has been 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 set to be 1), m is the dimension of the input space, and σ^2 is the variance of the data.[40] It means that the value of c depends on the system under the study. The five descriptors in the two latent variables space chosen by the GA-KPLS feature selection methods which were contained. These descriptors were obtained constitutional descriptors (number of bonds (nBT)), charge descriptors (submolecular polarity parameter (SPP)), molecular properties (topological polar surface area using N, O, S, P polar contributions (TPSA (Tot)) and Moriguchi octanol-water partition coeff. (logP) (MLOGP)) and quantum chemical descriptors (LUMO). For the constructed model, four general statistical parameters were selected to evaluate the prediction ability of the model for the log K_s. The statistical parameters Q^2, RE, AbsE, and RMSE were obtained for proposed models. Each of the statistical parameters mentioned were used for assessing the statistical significance of the QSPR model. Inspection of the results reveals a higher Q^2 and lowers other values parameter for the training and test sets GA-KPLS compared with their counterparts for GA-PLS. The GA-PLS linear model has good statistical quality

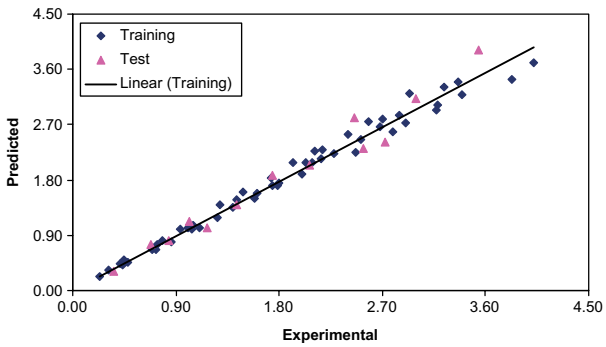

Figure 1. Plot of predicted log K_s obtained by GA-KPLS against the experimental values.

with low prediction error, while the corresponding errors obtained by the GA-KPLS model are lower. Plots of predicted log K_s versus experimental log K_s values by GA-KPLS for training and test set are shown in Figure 1. 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 to one and zero. This clearly shows the strength of GA-KPLS as a non-linear feature selection method. This result indicates that the log K_s of drug molecules possesses some non-linear characteristics.

Model validation

Validation is a crucial aspect of any QSPR/QSRR modelling.[41] The accuracy of 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 are created by deleting, in each case, one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modelling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones or group data that have not been utilized in the development of the model).[42] In particular, the LGO 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. 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 following equation:

$$R^2_{cv} \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^{\wedge})^2}{\sum_{i=1}^{n}(y_i - y^{-})^2} \quad (9)$$

where y_i, $y_i^{\wedge}$ and 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 ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model.[43] However, this assumption is, in many cases, incorrect and can be that 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.[41] Thus, the high value of Q^2 appears to be a necessary but not 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 is estimated on the predicted values of test set chemicals. The data set was randomly divided into two groups including training set (calibration and prediction sets) and test set, which consists of 52 and 13 molecules, respectively. The calibration set was used for model generation. The prediction set was applied to deal with over fitting of the network, whereas the test set (its molecules have no role in model building) was used for the evaluation of the predictive ability of the models for the 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 this analysis, the descriptive power of a given model has been 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

The liposomes possess an orderly molecular structure and are able to exert electrostatic interactions.[44,45] 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 negatively charged EPL liposome increased the retention of positively charged drugs and decreased retention of negatively charged drugs. Drugs with low log *K*s values had higher retention on the EPL liposome.[31]

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 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.[46]

Constitutional descriptors are the simplest and most commonly used, 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 bounds, absolute and relative numbers of specific atom type, absolute and relative numbers of single, double, triple, and aromatic bounds, number of rings, number of rings divided by the number of atoms or bonds, number of benzene rings, number of benzene rings divided by the number of atoms, molecular weight and average molecular weight.

The weighted holistic invariant molecular (WHIM) descriptors are built in such a way as to capture the relevant molecular 3-D information regarding the molecular size, shape, symmetry, and atom distribution with respect to some invariant reference frame. These descriptors are quickly computed from the atomic positions of the molecule atoms (hydrogens included). WHIM descriptors are based on principal component analysis of the weighted covariance matrix obtained from the atomic Cartesian coordinates. In relation to the kind of weights selected for the atoms different sets of WHIM descriptors can be obtained. Unitary weights (u), atomic mass (m), atomic van der Waals volume (v), atomic electronegativity (e), atomic polarizability (p), and atomic electrotopological state (s) are the available weighting schemes globally providing 66 directional and 33 global WHIM descriptors.

Although lipophilicity, molecular volume, WHIM 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 atoms, bonds, and molecular fragments. Electronic properties may play a role in the magnitude in a biological activity, along with structural features encoded in indexes. The eigenvalues 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 likely to be less 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 were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such 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.[47]

From this discussion, it can be seen that the molecular size, hydrogen bond, and electrostatic interactions are the likely three 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 this research, an accurate QSPR model for estimating the partitioning of drugs (log K_s) was developed by employing the one linear model (GA-PLS) and one non-linear model (GA-KPLS). 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. Two models have good predictive capacity and excellent statistical parameters. A comparison between these models revealed the superiority of the GA-KPLS to the GA-PLS model. It is easy to notice that there was a good prospect for the GA-KPLS application in the QSPR modelling. This indicates that log K_s of drug molecules possesses some non-linear characteristics. The results showed that the GA-KPLS model can be effectively used to describe the molecular structure characteristic of these compounds. It can also be used successfully to estimate the log K_s for new compounds or for other compounds whose experimental values are unknown.

References

[1] Q. Yang, P. Lundahl. Steric immobilization of liposomes in chromatographic gel beads and incorporation of integral membrane proteins into their lipid bilayers. *Anal. Biochem.* **1994**, *218*, 210.

[2] Y. Zhang, C.-M. Zeng, Y.-M. Li, S. Hjerten, P. Lundahl. Immobilized liposome chromatography of drugs on capillary continuous beds for model analysis of drug-membrane interactions. *J. Chromatogr. A* **1996**, *749*, 13.

[3] Q. Yang, X.-Y. Liu, S.-I. Ajiki, M. Hara, P. Lundahl, J. Miyake. Avidin–biotin immobilization of unilamellar liposomes in gel beads for chromatographic analysis of drug–membrane partitioning. *J. Chromatogr. B* **1998**, *707*, 131.

[4] Q. Yang, X. Y. Liu, M. Yoshimoto, R. Kuboi, J. Miyake. Covalent immobilization of unilamellar liposomes in gel beads for chromatography. *Anal. Biochem.* **1999**, *268*, 354.

[5] Y. Wang, L. Kong, X. Lei, L. Hu, H. Zou, E. W. beck, S.W. Annie Bligh, Z. Wang. Comprehensive two-dimensional high-performance liquid chromatography system with immobilized liposome chromatography column and reversed-phase column for separation of complex traditional Chinese medicine Longdan Xiegan Decoction. *J. Chromatogr. A* **2009**, *1216*, 2185.

[6] C. Huang, J. T. Mason. Geometric packing constraints in egg phosphatidylcholine vesicles. *P. Natl Acad. Sci. USA.* **1978**, *75*, 308.

[7] S. Ong, H. Liu, C. Pidgeon. Immobilized-artificial-membrane chromatography: measurements of membrane partition coefficient and predicting drug membrane permeability. *J. Chromatogr. A* **1996**, *728*, 113.

[8] P. Lundahl, F. Beigi. Immobilized liposome chromatography of drugs for model analysis of drug-membrane interactions. *Adv. Drug Deliv. Rev.* **1997**, *23*, 221.

[9] X. Liu, Q. Yang, N. Kamo, J. J. Miyake. Effect of liposome type and membrane fluidity on drug–membrane partitioning analyzed by immobilized liposome chromatography. *J. Chromatogr. A* **2001**, *913*, 123.

[10] C. Pidgeon, S. Ong, H. Chol, H. Liu. Preparation of mixed ligand immobilized artificial membranes for predicting drug binding to membranes. *Anal. Chem.* **1994**, *66*, 2701.

[11] C. Pidgeon, S. Ong. Predicting drug-membrane interactions. *Chemtech.* **1995**, *25*, 38.

[12] S. Ong, H. Liu, X. Qiu, C. Pidgeon. Membrane partition coefficients chromatographically measured using immobilized artificial membrane surfaces. *Anal. Chem.* **1995**, *67*, 755.

[13] C. Y. Yang, S. J. Cai, H. Liu, C. Pidgeon. Immobilized artificial membranes — screens for drug membrane interactions. *Adv. Drug Deliv. Rev.* **1996**, *23*, 229.

[14] P. Artursson, J. Karlsson. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochem. Biophys. Res. Commun.* **1991**, *175*, 880.

[15] C. Altomare, R. Tsai, N. E. Tayar, B. Testa, A. Carotti, S. Cellamare, P. G. D. Benedetti. Determination of lipophilicity and hydrogen-bond donor acidity of bioactive sulphonyl-containing compounds by reversed-phase HPLC and centrifugal partition chromatography and their application to structure-activity relations. *J. Pharm. Pharmacol.* **1991**, *43*, 191.

[16] P. Artursson. Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorbtive (Caco-2) cells. *J. Pharm. Sci.* **1990**, *79*, 476.

[17] F. Beigi, I. Gottschalk, C. Lagerquist Hägglund, L. Haneskog, E. Brekkan, Y. Zhang, T. Österberg, P. Lundahl. Immobilized liposome and biomembrane partitioning chromatography of drugs for prediction of drug transport. *J. Pharm.* **1998**, *164*, 129.

[18] C. Lagerquist, F. Beigi, A. Karlén, H. Lennernäs, P. Lundahl. Effects of cholesterol and model transmembrane proteins on drug partitioning into lipid bilayers as analysed by immobilized-liposome chromatography. *J. Pharm. Pharmacol.* **2001**, *53*, 1477.

[19] E. Boija, A. Lundquist, J. J. Martínez Pla, C. Engvall, P. Lundahl. Effects of ions and detergents in drug partition chromatography on liposomes. *J. Chromatogr. A* **2004**, *1030*, 273.

[20] N. Yoshimoto, M. Yoshimoto, K. Yasuhara, T. Shimanouchi, H. Umakoshi, R. Kuboi. Evaluation of temperature and guanidine hydrochloride-induced protein-liposome interactions by using immobilized liposome chromatography. *Biochem. Eng. J.* **2006**, *29*, 174.

[21] G. Klopman, H. Zhu. Recent methodologies for the estimation of n-octanol/water partition coefficients and their use in the prediction of membrane transport properties of drugs. *Mini Rev. Med. Chem.* **2005**, *5*, 127.

[22] G. Schuurmann, R. U. Ebert, R. Kuhne. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. *Environ. Sci. Technol.* **2006**, *40*, 7005.

[23] X. Binbin, M. Weiping, Z. H. Xiaoyun, F. Botaon. Quantitative structure–retention relationships for organic pollutants in biopartitioning micellar chromatography. *Anal. Chim. Acta* **2007**, *598*, 12.

[24] J. M. Bermudez-Saldana, L. Escuder-Gilabert, M. J. Medina-Hernandez, R.M. Villanueva-Camanas, S. Sagrado. Chromatographic retention–activity relationships for prediction of the toxicity pH-dependence of phenols. *Chemosphere* **2007**, *69*, 108.

[25] M. Weiping, L. Feng, Z. H. Haixia, Z. H. Xiaoyun, L. Mancang, H. Zhide, F. Botao. Quantitative structure–property relationships for pesticides in biopartitioning micellar chromatography. *J. Chromatogr. A* **2006**, *1113*, 140.

[26] A. Gajewicz, M. Haranczyk, T. Puzyn. Predicting logarithmic values of the subcooled liquid vapor pressure of halogenated persistent organic pollutants with QSPR: How different are chlorinated and brominated congeners? *Atmos. Environ.* **2010**, *44*, 1428.

[27] H. Noorizadeh, A. Farmany. QSRR models to predict retention indices of cyclic compounds of essential oils. *Chromatographia* **2010**, *72*, 563.

[28] H. Noorizadeh, A. Farmany, A. Khosravi. Investigation of retention behaviors of essential oils by using QSRR. *J. Chin. Chem. Soc.* **2010**, *57*, 1.

[29] N. Krämer, A. L. Boulesteix, G. Tutz. Penalized partial least squares with applications to B-spline transformations and functional data. *Chemom. Intell. Lab. Syst.* **2008**, *94*, 60.

[30] S. Haykin, *Neural Networks*, Prentice-Hall: New Jersey, **1999**.

[31] T. Osterberg, M. Svensson, P. Lundahl. Chromatographic retention of drug molecules on immobilised liposomes prepared from egg phospholipids and from chemically pure phospholipids. *Eur. J. Pharm. Sci.* **2001**, *12*, 427.

[32] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, *DRAGON-Software for the Calculation of Molecular Descriptors*. Version 3.0 for Windows, **2003**.

[33] D. E. Goldberg, Genetic algorithms, in *Search, Optimization and Machine Learning*, Addison-Wesley – Longman: Reading, MA, USA, **2000**.

[34] S. Riahi, E. Pourbasheer, R. Dinarvand, M. R. Ganjali, P. Norouzi. Exploring QSARs for antiviral activity of 4-Alkylamino-6-(2-hydroxyethyl)-2-methylthiopyrimidines by support vector machine. *Chem. Biol. Drug Des.* **2008**, *72*, 205.

[35] H. Noorizadeh, A. Farmany, *J. Chin. Chem. Soc.* **2010**, *57*, 1.

[36] U. Depczynski, V. J. Frost, K. Molt. Genetic algorithms applied to the selection of factors in principal component regression. *Anal. Chim. Acta* **2000**, *420*, 217.

[37] S. Wold, M. Sjostrom, L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109.

[38] B. M. Nicolaï, K. I. Theron, J. Lammertyn. Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 243.

[39] R. Rosipal, L. J. Trejo, *J. Mach. Learn. Res.* **2001**, *2*, 97.

[40] K. Kim, J. M. Lee, I. B. Lee. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. *Chemom. Intell. Lab. Syst.* **2005**, *79*, 22.

[41] J. Acevedo-Martınez, J. C. Escalona-Arranz, A. Villar-Rojas, F. Tellez-Palmero, R. Perez-Roses, L. Gonzalez, R. Carrasco-Velar. Quantitative study of the structure–retention index relationship in the imine family. *J. Chromatogr. A* **2006**, *1102*, 238.

[42] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou. A novel QSAR model for predicting induction of apoptosis by 4-aryl-4*H*-chromenes. *Bioorg. Med. Chem.* **2006**, *14*, 6686.

[43] A. Golbraikh, A. Tropsha. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269.

[44] M. M. Felix, H. Umakoshi, T. Shimanouchi, M. Yoshimoto, R. Kuboi. Evaluation of interaction between liposome membranes induced by stimuli responsive polymer and protein. *J. Biosci. Bioeng.* **2002**, *93*, 498.

[45] E. Boija, A. Lundquist, J. J. Martínez Pla, C. Engvall, P. Lundahl. Effects of ions and detergents in drug partition chromatography on liposomes. *J. Chromatogr. A* **2004**, *1030*, 273.

[46] J. V. Turner, B. D. Glass, S. Agatonovic-Kustrin. Prediction of drug bioavailability based on molecular structure. *Anal. Chim. Acta* **2003**, *485*, 89.

[47] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, Germany, **2000**.