Journal of Computer-Aided Molecular Design (2005) 19: 499–508
DOI 10.1007/s10822-005-9003-5

Prediction of the tissue/blood partition coefficients of organic compounds based on the molecular structure using least-squares support vector machines

H.X. Liu[a], X.J. Yao[a], R.S. Zhang[b], M.C. Liu[a], Z.D. Hu[a,]* & B.T. Fan[c]
[a]*Department of Chemistry, Lanzhou University, 730000, Lanzhou, China;* [b]*Department of Computer Science, Lanzhou University, 730000, Lanzhou, China;* [c]*Université Paris 7-Denis Diderot, ITODYS 1 rue Guy de la Brosse, 75005, Paris, France*

Received 27 March 2005; accepted 6 July 2005

Key words: heuristic method, least-squares support vector machines (LS-SVM), QSAR, tissue/blood partition coefficients

Summary

The accurate nonlinear model for predicting the tissue/blood partition coefficients (PC) of organic compounds in different tissues was firstly developed based on least-squares support vector machines (LS-SVM), as a novel machine learning technique, by using the compounds' molecular descriptors calculated from the structure alone and the composition features of tissues. The heuristic method (HM) was used to select the appropriate molecular descriptors and build the linear model. The prediction result of the LS-SVM model is much better than that obtained by HM method and the prediction values of tissue/blood partition coefficients based on the LS-SVM model are in good agreement with the experimental values, which proved that nonlinear model can simulate the relationship between the structural descriptors, the tissue composition and the tissue/blood partition coefficients more accurately as well as LS-SVM was a powerful and promising tool in the prediction of the tissue/blood partition behaviour of compounds. Furthermore, this paper provided a new and effective method for predicting the tissue/blood partition behaviour of the compounds in the different tissues from their structures and gave some insight into structural features related to the partition process of the organic compounds in different tissues.

Introduction

Drug discovery and development are expensive undertakings. The research costs for a compound increase dramatically as it progresses through clinical development, and therefore there are economic reasons for identifying and discontinuing the development of poor drug candidates at the earliest possible time. In order to reduce the cost of drug discovery/development, *in silico* approaches are being used today in drug discovery. Among the *in silico* approaches, physiologically based pharmacokinetic (PBPK) modeling is a promising tool, because – in contrast to classical compartmental analysis – PBPK modeling attempts to describe the system in physiological terms that have relevance to chemical distribution, mode of action and underlying biochemical processes. In essence, PBPK modeling offers a scientifically-defensible way – instead of an educated guess – to integrate various pieces of information from *in silico* models, *in vitro* studies and other preclinical information, to evaluate the outcome under various assumptions. Furthermore, PBPK modeling is an excellent tool for simulating variability at different levels, organ, organism (inter-individual) and population (interethnic) [1].

*To whom correspondence should be addressed. Fax: +86-931-891-2582, E-mail: huzd@lzu.edu.cn

In a word, PBPK models play an important role and are being increasingly used as a tool in drug discovery/development. The form of the models depends predominantly on the rate of the tissue/blood distribution of the compounds. Thus, for PBPK modeling, the tissue/blood partition coefficients (PC) of the drug in various organs and tissues need to be known. However, the experimental determination of the tissue/blood partition coefficients is difficult, expensive, and time-consuming, because it involves the direct measurement of the drug concentration in the tissue/blood of laboratory animals, and obviously requires the synthesis of the compounds, often in the radiolabeled form [2, 3]. Therefore, it should be extremely useful to predict the tissue/blood partition coefficients of compounds.

Importance of the tissue/blood partition coefficients in pharmacokinetic modeling for risk assessment and drug design led several investigators to analyze the factors influencing the tissue/blood partition coefficients. Zhang showed that the tissue/blood partition coefficients of a compound depend strongly on tissue compositions [3, 4]. Baláž et al. described the tissue/blood partition coefficients as a nonlinear function of lipophilicity and tissue composition [2]. However, in these models, molecular structure of the object compounds was not considered enough and then the accuracy of prediction was not high.

Quantitative structure-activity relationship (QSAR) provides a promising method for the estimation of tissue/blood partition of compounds based on the descriptors derived solely from the molecular structure to fit experimental data, which has been successfully established to predict different important biopharmaceutical properties, such as metabolism [5], toxicity [6], oral bioavailability [7], etc. The advantage of this approach over other methods lies in the fact that it requires only the knowledge of chemical structure and is not dependent on the experiment properties. This study can develop a method for the prediction of the property of new compounds that have not been synthesized or found. It can also identify and describe important structural features of the molecules that are relevant to variations in molecular properties, thus, gain some insight into the structural factors affecting the molecular properties. Computational models of this type are useful because they rationalize a large number of experimental observations and therefore allow save time and money in the drug design process. In addition, they are useful in areas like design of virtual compound libraries, computational-chemical optimization of compounds, and design of combinatorial libraries with appropriate ADME (absorption, distribution, metabolism and excretion) properties and development of PBPK models.

Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications [8–16] SVM shows outstanding performances because it can lead to global models that are often unique by embodies the Structural Risk Minimization (SRM) principle [17, 18], which has been shown to be superior to the traditional Empirical Risk Minimization (ERM) principle. Furthermore, due to their specific formulation, sparse solutions can be found, and both linear and nonlinear regression can be performed. However, finding the final SVM model can be computationally very difficult because it requires the solution of a set of nonlinear equations (quadratic programming). As a simplification of "traditional" SVM, Suykens and Vandewalle [19] have proposed the use of Least-Squares SVM (LS-SVM). LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple.

Another important factor responsible for the quality of the QSPR/QSAR model is the numerical representation (often called molecular descriptor) of the chemical structure. The performance and the accuracy of the results are strongly dependent on the way the structures are represented. The Software CODESSA, developed by Katritzky group, enables the calculation of a large number of quantitative descriptors based solely on the molecular structural information and codes chemical information into mathematical form [20, 21]. CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish molecular structure-property/activity relationships. CODESSA has been applied successfully in a variety of QSAR analyses [22, 23].

In the present investigation, LS-SVM, as a novel machine learning technique, for the first time, was used for the prediction of the tissue/blood partition behavior of organic compounds using calculated molecular descriptors based on the software CODESSA and the weight fractions of water, protein in different tissues as inputs. The aim was to explore the tissue/blood partition behavior of organic compounds in different tissues (the kidney, brain, muscle, lung, liver, heart, and fat), to develop an accurate quantitative model correlating the structural descriptors, tissue composition and the tissue/blood partition coefficients, and at the same time, to seek for the structural factor affecting the tissue/blood partition behavior of organic compounds and the essential differences of the different tissues. The prediction results were very satisfactory in both training set and test set compounds, which proved LS-SVM a powerful and useful tool.

Methodology

Data preparation

The tissue/blood partition coefficients (expressed as the logarithm of the partition coefficient, logPC) of 35 organic chemicals for human fat, liver, brain, kidney, muscle, lung, and the heart were taken from [2] and are listed in Table 1. Among the 245 (35*7) samples, there are 37 samples without the tissue/blood partition coefficient values (blank in the table). The remained 208 samples were used to build model. The data set was randomly divided into a training set of 156 samples and a test set of 52 samples.

Calculation of the descriptors

To obtain a QSPR/QSAR model, compounds are often represented by the molecular descriptors. The calculation process of the molecular descriptors is described as below: All molecules were drawn into Hyperchem and pre-optimized using MM+ molecular mechanics force field [24]. A more precise optimization was done with semi-empirical AM1 method in MOPAC [25]. All calculations were carried out at restricted Hartree Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.01. The MOPAC output files were used by the CODESSA program to calculate five classes of descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier–Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.) [20].

In order to predict the tissue/blood partition of 35 organic compounds in different tissues simultaneously and consider the influences of the tissues on the tissue/blood partition of these compounds, the weight fractions of water, protein, lipid in different tissues, taken from the reference [2], were added into the descriptors tool to represent the properties of different tissues. They were given in Table 2.

Selection of descriptors based on the heuristic method [20]

Once molecular descriptors are generated, the heuristic method in CODESSA was used to accomplish the pre-selection of the descriptors and build the linear model. Its advantages are the high speed and no software restrictions on the size of the data set. The heuristic method can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which descriptors are insignificant (from the standpoint of a single-parameter correlation), and which descriptors are highly intercorrelated. This information will be helpful in reducing the number of descriptors involved in the search for the best QSAR/QSPR model.

First of all, all descriptors are checked to ensure: (a) that values of each descriptor are available for each structure and (b) that there is a variation in these values. Descriptors for which values are not available for every structure in the data in question are discarded. Descriptors having

Table 1. The experimental and predicted tissue/blood partition coefficients (logPC) by LS-SVM.

No	Compounds	Kidney		Brain		Muscle		Lung		Liver		Heart		Fat	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
1	2,2-Dimethylbutane	0.73	0.61	1.03	0.95	0.58	0.56	0.36	0.58	1.13	0.73	0.28	0.71	2.4	2.41
2	Pentane	0.20	0.29	0.76	0.49	0.27	0.26	0.12	0.29	0.74	0.40	−0.28	0.36	2.02	1.93
3	2-Me–Pentane	0.69	0.63	0.97	0.91	0.85	0.59	0.29	0.62	1.04	0.75	0.53	0.71	2.33	2.32
4	3-Me–Pentane	0.76	0.65	1.01	0.91	0.95	0.60	0.32	0.63	1.06	0.76	0.65	0.73	2.38	2.27
5	Hexane	0.57	0.45	0.80	0.70	0.80	0.41	0.10	0.44	0.81	0.56	0.54	0.53	2.11	2.11
6	3-Me–Hexane	0.75	0.62	0.89	0.91	0.93	0.57			0.93	0.75	0.61	0.72	2.33	2.31
7	Heptane	0.67	0.58	0.80	0.87	0.80	0.54	0.12	0.57	0.76	0.70	0.51	0.67	2.31	2.26
8	Cyclopropane	0.15	0.11	0.08	0.12	0.12	0.09			0.19	0.21			1.45	1.43
9	Me–Cyclopentane	0.74	0.65	0.93	0.94	0.76	0.61	0.3	0.63	0.96	0.78	0.34	0.76	2.31	2.34
10	Cyclohexane	0.74	0.69	0.93	0.96	0.89	0.64	0.32	0.67	0.93	0.81	0.65	0.78	2.3	2.32
11	Benzene	0.27	0.26	0.45	0.42	0.4	0.23	0.27	0.26	0.56	0.37	0.34	0.32	1.82	1.87
12	Toluene	0.24	0.43	0.55	0.67	0.53	0.39	0.31	0.42	0.67	0.54	0.46	0.50	1.97	2.09
13	Styrene									0.43	0.56			1.7	2.10
14	CH_2Cl_2	−0.01	−0.04	0	0.09	−0.1	−0.04	−0.01	−0.01	0.08	0.01	−0.14	−0.03	1.15	1.06
15	$CHCl_3$	0.14	0.11	0.4	0.30	0.18	0.09	−0.06	0.12	0.33	0.20	−0.11	0.16	1.54	1.56
16	Me–CCl_3	0.31	0.35	0.4	0.46	0.31	0.30	0.15	0.34	0.69	0.48	0.44	0.43	1.88	1.88
17	CF_3–CH_2Cl	−0.16	0.06	0.08	0.14	0.17	0.04	0.17	0.08	0.12	0.14			1.36	1.33
18	Teflurane			0.26	0.33	0.58	0.25			0.22	0.39			1.52	1.71
19	Halothane	0.41	0.45	0.39	0.44	0.54	0.40	0.24	0.45	0.57	0.62			1.97	2.05
20	$CF_2=CHCl$	0.06	−0.00	0.1	0.06	0.06	−0.01	0.06	0.02	0.1	0.05			1.36	1.02
21	$CCl_2=CHCl$	0.27	0.29	0.41	0.35	0.37	0.27	0.24	0.31	0.55	0.41	0.32	0.35	1.85	1.73
22	MeOH	−0.08	−0.11	−0.17	−0.28	−0.09	−0.07	0.03	−0.04					−0.85	−0.64
23	EtOH	−0.15	−0.17	−0.23	−0.32	−0.19	−0.14	−0.06	−0.11					−0.8	−0.79
24	1-PrOH	−0.14	−0.17	−0.22	−0.25	−0.15	−0.13	−0.07	−0.11					−0.51	−0.58
25	2-PrOH	−0.16	−0.18	−0.22	−0.23	−0.16	−0.15	−0.09	−0.13					−0.6	−0.48
26	1-BuOH	−0.16	−0.14	−0.21	−0.18	−0.2	−0.12	−0.13	−0.10					−0.14	−0.36
27	Acetone	−0.13	−0.16	−0.21	−0.19	−0.11	−0.14	−0.09	−0.11					−0.36	−0.22
28	Butanone	−0.07	−0.10	−0.11	−0.10	−0.09	−0.08	−0.09	−0.06					0.11	0.01
29	Et_2O	−0.04	−0.04	0.03	0.04	−0.08	−0.05			−0.04	0.03	0	−0.02	0.77	1.16
30	Divinyl ether	−0.17	−0.03	0.16	0.02	−0.04	−0.03			0.05	0.03			1.2	0.98
31	Fluroxene	−0.03	0.12	0.16	0.23	0.16	0.11			0.16	0.20			1.39	1.44
32	Enflurane	0.28	0.28	0.26	0.33	0.43	0.25	0.3	0.29	0.45	0.39			1.87	1.71
33	Isoflurane	0.18	0.21	0.32	0.25	0.23	0.18	0.06	0.22	0.47	0.33			1.69	1.64
34	Methoxyflurane	0.18	0.25	0.29	0.25	0.29	0.21	0.06	0.26	0.4	0.40			1.8	1.75
35	Sevoflurane	0.52	0.32	0.34	0.36	0.4	0.28	0.34	0.33	0.7	0.44			1.86	1.78

a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested and insignificant descriptors removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. All two-parameter regression models with remaining descriptors are subsequently developed and ranked by the regression correlation coefficient R^2. A stepwise addition of further descriptor scales is performed to find the best multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, the crossvalidated R^2_{cv}, and the F-value).

The heuristic method usually produces correlations 2–5 times faster than other methods with comparable quality [26]. The rapidity of calculations from the heuristic method renders it the first method of choice in practical research. Thus, in the present investigation, we used this method to select structural descriptors and build the linear model.

Table 2. Tissue composition (weight fractions).

Tissue	Lipid	Protein	Water
Fat	0.8000	0.050	0.150
Liver	0.0700	0.180	0.720
Brain	0.1070	0.079	0.790
Kidney	0.0500	0.170	0.770
Muscle	0.0200	0.170	0.790
Lung	0.0100	0.177	0.780
Heart	0.1000	0.167	0.727

Least squares support vector machine

In recent years, the support vector machine (SVM), as a powerful new tool for data classification and function estimation, has been developed [27]. SVM maps input data into a high dimensional feature space where it may become linearly separable. Recently SVM has been applied to a wide variety of domains such as pattern recognition and object detection [17], function estimation [28] etc. One reason that SVM often performs better than earlier methods is that SVM was designed to minimize structural risk whereas previous techniques are usually based on minimization of empirical risk. So SVM is usually less vulnerable to overfitting problem. Especially, Suykens and Vandewalle [19] proposed a modified version of SVM called least squares SVM (LS-SVM), which resulted in a set of linear equations instead of a quadratic programming problem, which can extend the application of the SVM.

There exist a number of excellent introductions for SVM, both printed [28–30] and electronically available [31]. The theory of LS-SVM for classification and function estimation has also been described clearly by Suykens and Vandewalle [19]. For this reason, we only briefly described the differences between SVM and LS-SVM for function estimation here.

In principle, LS-SVM always fits a linear relation ($y = wx + b$) between the regressors (x) and the dependent variable (y). The best relation is the one that minimizes the cost function (Q) containing a penalized regression error term:

$$Q_{\text{LS-SVM}} = \frac{1}{2} w^T w + \gamma \sum_{k=1}^{N} e_k^2 \qquad (1)$$

subject to

$$y_k = w^T \varphi(x_k) + b + e_k, \quad k = 1, \ldots, N$$

The first part of this cost function is a so-called L_2 norm on the regression weights. Using this norm, weight values are penalized quadratically, and it aims at coefficients that are as small as possible. The second term takes into account the regression error (e_k) for all of the N training objects (the standard Least-Squares error approach). The relative weight of this part as compared to the first part is indicated by the parameter γ, which has to be optimized by the user. The third part gives the definition of the regression error to be the difference between the true and predicted values, and this can be seen as a constraint. For comparison, note that the traditional SVM approach defines the regression error differently by neglecting all regression errors smaller than $\pm\epsilon$ (the ϵ-insensitive loss function). It is this difference in error definitions that makes the LS-SVM optimization problem computationally much easier than the original SVM problem. Furthermore, the value of parameter ϵ does not have to be optimized for LS-SVM, which is the case for SVMs.

Similar to SVM, the LS-SVM also considers this optimization problem to be a constrained optimization problem and uses a Lagrange function to solve it. By solving the Lagrange style of Equation 1, the weight coefficients (w) can be written as an expansion of Lagrange multipliers with the corresponding training objects:

$$w = \sum_{k=1}^{N} \alpha_k x_k \text{ with } \alpha_k = 2\gamma e_k \qquad (2)$$

By substituting (2) into the original regression line ($y = wx + b$), the following result can be obtained:

$$y = \sum_{k=1}^{N} \alpha_k x_k^T x + b \qquad (3)$$

It can be seen that the Lagrange multipliers can be defined as

$$\alpha_k = (x_k^T x + (2\gamma)^{-1})^{-1}(y_k - b) \qquad (4)$$

Finding these Lagrange multipliers is very simple as opposed to the SVM approach in which a more difficult relation has to be solved to obtain these values.

In addition, due to solving the optimization problem in terms of Lagrange multipliers like

SVM, LS-SVM has the same advantage that the final model can be written as a weighted linear combination of the inner product between the training points and a new test object. Therefore, it easily allows nonlinear regression as an extension of the linear approach by introducing the kernel function. This leads to the following nonlinear regression function:

$$f(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \qquad (5)$$

In Equation 5, $K(x, x_k)$ is the kernel function. The value is equal to the inner product of two vectors x and x_k in the feature space $\Phi(x)$ and $\Phi(x_k)$, that is, $K(x,x_k) = \Phi(x)^T \bullet \Phi(x_k)$. The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function. The Gaussian kernel $K(x,x_k) = \exp(-\|x_k - x\|^2/\sigma^2)$ is commonly used.

Note that, in contrast to the Lagrange multipliers, the choice of a kernel and its specific parameters together with γ do not follow from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik–Chervonenkis bounds, cross validation, an independent optimization set, or Bayesian learning. In this paper, the Gaussian kernel was used as kernel function and the 10 fold cross-validation was used to tune the optimized values of the two parameter γ and σ. The MSE was used as the error function, and it is computed according to the following equation

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (d_i - o_i)^2}{n}$$

where d_i are the teaching outputs (desired outputs), o_i are the actual outputs, and n is the number of samples.

All calculations of LS-SVM was performed using the Matlab/C toolbox [32].

Results and discussion

Results of the heuristic method

In order to select the descriptors responsible for the tissue/blood partitioning, the heuristic method was used. About 500 descriptors were calculated by the CODESSA program for each of the compounds. Besides these structural descriptors of compounds, three descriptors describing the composition of tissue the weight fractions of protein, lipid, water in different tissues, were also added into the pool of descriptors. After the heuristic reduction, the pool of descriptors was reduced to 147. A variety of subset sizes were investigated to determine the optimum number of descriptors in a model. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. The influences of the number of the descriptors on the correlation coefficient (R^2) and the standard deviation (s^2) were shown in Figure 1. According to Figure 1, the five-parameter model was chosen as the best linear model. The statistical analysis results of the five-parameter model and the involved molecular descriptors as well as their corresponding physical–chemical meaning were summarized in Table 3.

By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the tissue/blood partition coefficient of organic compounds. The tissue/blood partition coefficient is defined as the ratio of the equilibrium concentrations of the compound in the tissue and in blood. In both phases, the compound is present as accumulated in lipids, bound to proteins, and dissolved in the interstitial and intracellular aqueous phases. For a well-absorbed drug, the distribution in tissue is affected by the interactions between the drug and lipid, protein. Lipid mainly interacts with lipophilic molecules, and protein mainly interacts with hydrophilic and charged molecules.

In the linear model, there is one tissue composition descriptor, two electrostatic descriptors, one quantum chemical descriptor and one geometrical descriptor. According to the t-test (Table 3), the most important descriptor affecting the tissue/blood partition coefficient is a parameter describing the composition of tissue, weight fractions of water in different tissues. It is apparent and easy to understand the importance of tissue composition for the tissue/blood partition. The negative coefficient indicates that the larger the weight fractions of water in tissues, the smaller the tissue/blood partition coefficient, which is in agreement with the reference [2].

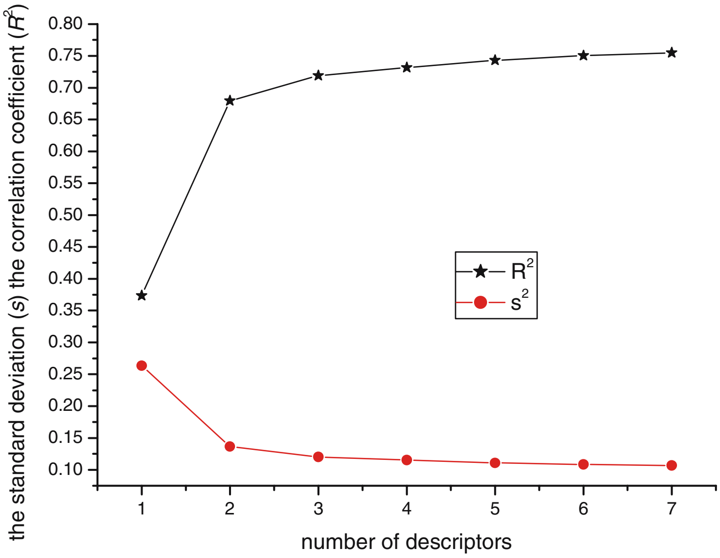

Figure 1. Influence of the number of descriptors on the correlation coefficient (R^2) and the standard deviation (s^2) of the regression models.

Both the two electrostatic descriptors are CPSA (Charged Partial Surface Area) descriptors *RNCG Relative negative charge* and *WPSA-1 Weighted PPSA* (PPSA-1*TMSA/1000), where PPSA1 is partial positive charge surface and TMSA is total molecular surface area [33]. These two descriptors describe the negative and positive partial charge distribution information in the molecule and then can account for the electrostatic interaction between the compound and the protein.

The quantum chemical descriptor $N_{\text{H-donors,}}$ count of H-donors sites, reflects the hydrophilicity of a compound. Generally, hydrophilicity unfavors compounds into lipid and favors compounds into water. Furthermore, H-donors sites of compounds lead to reduced lipid absorption.

The last descriptor is a geometric descriptor, YZ Shadow. It can be defined as the area of shadows of the molecule as projected on the YZ plane by the orientation of the molecule in the space along the axes of inertia, which characterizes the size and geometrical shape of the molecule [20]. Thus, it can act as a descriptor of Van der Waals and dispersion interactions between compounds and lipid. With the increasing of this descriptor, the Van der Waals and dispersion interactions between compounds and lipid interaction increase. This can lead to the increasing tendency to partition into lipid.

From the above discussion, the structural descriptors can account for the structural features responsible for the tissue/blood partition coefficients of the compounds and the parameter describing the composition of tissue can describe the main differences between the different tissues. From Table 3, it can be seen that the model of HM

Table 3. Descriptors, coefficients, standard error, and *t*-values for the linear model.[a]

Descriptor	Chemical meaning	Coefficient	Error	*t*-test
(Constant)	Intercept	1.2015	0.2068	5.8114
WF	Weight fractions of water in different tissues	−1.7503	−0.0994	−17.6171
RNCG	Relative negative charge (QMNEG/QTMINUS) [Zefirov's PC]	−0.5881	0.1688	−3.4852
WPSA-1	Weighted PPSA (PPSA1*TMSA/1000) [Quantum-Chemical PC]	0.0051	0.0008	6.3022
$N_{\text{H-donors}}$	Count of H-donors sites	−0.0533	0.0123	−4.3334
YZS	YZ Shadow	0.0224	0.0067	3.3596

[a] $R^2 = 0.74$; $s^2 = 0.11$; $n = 209$; $F = 117.42$; $R^2_{\text{cv}} = 0.72$.

was not sufficiently accurate ($R^2=0.74$; $s^2=0.11$), which could be caused by two factors: one side, the method itself, on the other hand, the factors influencing the tissue/blood partition behavior of the compounds were complex and not all of them were linear correlation with the tissue/blood partition coefficients. In addition, in this model, only one descriptor, the weight fractions of water in different tissues, was selected to differentiate the different tissues. However, for the tissues muscle and brain, this descriptor is completely same. Thus, this model will give the same predicted tissue/blood partition coefficients for the same compounds in the two tissues, which is not reasonable. In order to differentiate these two tissues, the weight fractions of protein or lipid in different tissues should occur for a reasonable model. The reason that this descriptor was not selected into the linear model was probably there is not apparent linear relationship between the tissue/blood partition coefficients and the weight fractions of protein or lipid. Furthermore, the partition coefficient between tissue and blood is an apparent partition coefficient between several compositions and blood. For a multiphase system, a better result can usually be obtained with the nonlinear model than with the linear equation. So, we built the nonlinear prediction models based on the selected descriptors together with the weight fractions of protein in different tissues by a novel method LS-SVM to further discuss the correlation between the molecular structure and the tissue/blood partition coefficients.

Result of LS-SVM

Optimizing LS-SVM

As discussed in the section 2.4, for LS-SVM, kernel function and its specific parameters together with γ have to be tuned by the user. In this paper, the radial basis function (RBF) kernel was used as kernel function. Thus γ (the relative weight of the regression error) and σ (the kernel parameter of the RBF kernel) need to optimized. Here, the optimal parameters are found from an intensive grid search. The result of this grid search is an error-surface spanned by the model parameters. A robust model is obtained by selecting those parameters that give the lowest error in a smooth area based on 10-fold cross validation of the training set.

To find the optimal model parameters and prevent the model from overfitting, a grid search is performed on the basis of 10-fold cross validation on the training set. The two parameters were tuned by a grid of 50*50, with the parameter (σ) of the RBF kernel, a wide range of σ^2 from 0.01–100 and the parameter γ, a range of 10–100,000. In this way, parameter optimization was performed in different orders of magnitude. Because the grid search has been performed over just two parameters, a contour plot of the optimization error can be visualized easily (Figure 2). This is an advantage of LS-SVMs over SVMs in which three parameters have to be optimized. From Figure 2, the optimal parameter settings can now be selected from a smooth subarea with a low cross validation MSE. The selected optimal value of σ^2 and γ is 37.28, 3.24×10^4, respectively, which point was marked by the red square in Figure 2.

The predicted result of LS-SVM

The predicted results of the optimal LS-SVM model ($\sigma^2=37.28$, $\gamma=3.24\times10^4$) were shown in Table 1 and Figure 3. As can be seen from Figure 2, the proposed models were statistically stable and fitted the data well. The mean square error of the training set, the test set and the whole set is 0.0226, 0.0289, 0.0242 and the prediction correlation coefficient is 0.970 and 0.974, 0.971 respectively. It can be concluded that the predicted values are in good agreement with the experimental values.

Analysis of the obtained results indicates that the model we proposed can correctly represent the relationship between the structure and partition coefficients of these compounds and that molecular descriptors calculated solely from structures could describe the structural features of the compounds responsible for their tissue/blood partition coefficients, at the same time, the weight fractions of water in different tissues could describe accurately the difference of the different tissues.

To test the suitability of the QSAR approach constructed by LSSVM, we have compared the obtained results with the results from the reference [2]. Table 4 showed the statistical parameters of the results obtained from the two studies for the same set of compounds. The RMS errors of the LSSVM model for whole

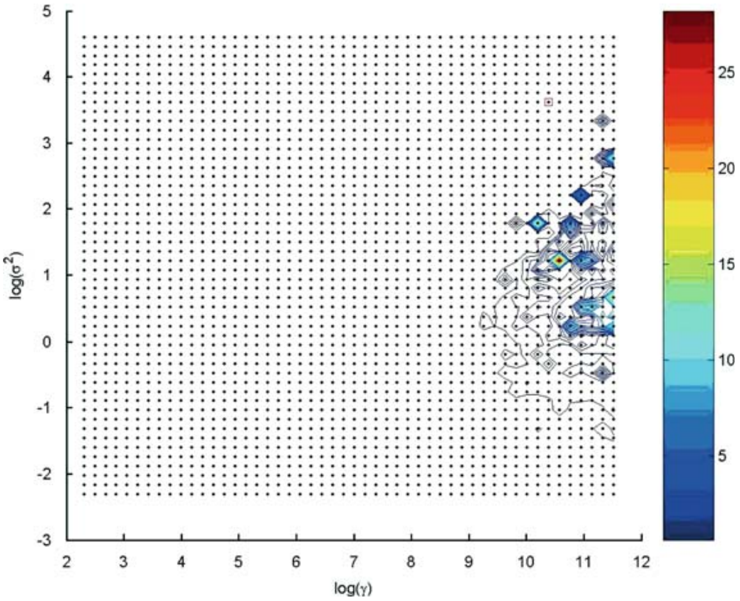

Figure 2. Contour plot of the optimization error for LS-SVM when optimizing the parameters σ and γ. The red square indicates the selected optimal settings.

data set were little lower than that of the model proposed in the reference [2]. The correlation coefficient (R^2) given by LSSVM model was higher than that of the models in the reference [2]. Through a regression analysis on the experimental and the calculated logPC obtained by different methods for the whole data set, the results of *F*-test and *t*-test were obtained and also shown in Table 4. From Table 4, it can be seen that the LSSVM model gives the little higher *F* and *t* values, so, this model gives the satisfactory results.

Conclusion

The least square support vector machine was used to develop the nonlinear model for predicting the tissue/blood partition coefficients of organic com-

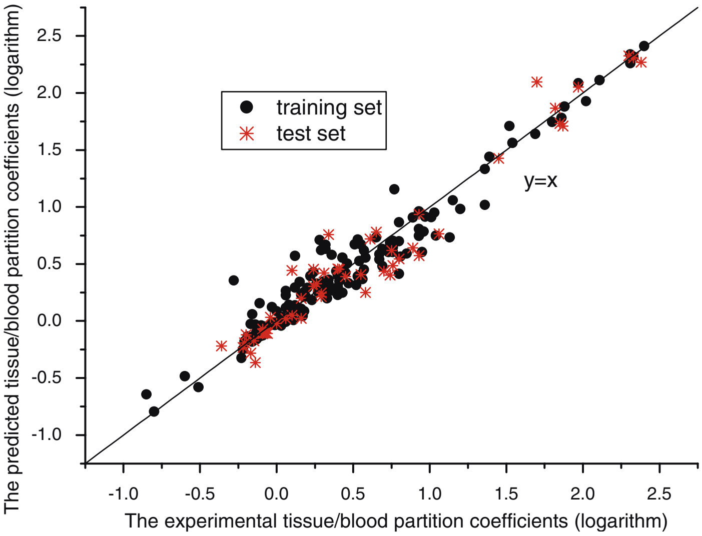

Figure 3. Plot of predicted log(PC) vs. experimental values for the training and test sets by LS-SVM.

Table 4. Comparisons of different QSPR models for the prediction of the logPC.

Models	RMS	R^2	*F*-test	*t*-test	Sig
Ref2	0.168	0.935	2971.806	54.514	0.000
LS-SVM	0.155	0.943	3376.736	58.110	0.000

pounds based on calculated descriptors of compounds and tissue composition for the first time. Very satisfactory results were obtained with the proposed method. By analyzing the obtained results, it can be concluded that: (1) The proposed models could identify and provide some insight into what structural features are related to the tissue/blood partition of the compounds. (2) The weight fractions of water in different tissues could describe accurately the difference of the different tissues. (3) Additionally, non-linear models using LS-SVM produced better models with good predictive ability than linear regression. LS-SVM proved to be a powerful and useful tool in the prediction of the pharmacokinetic property of organic compounds. LS-SVM has can lead to global (and often unique) nonlinear models and at the same time LS-SVM can be calculated easily and is easier to be controlled compared with SVM. Therefore, the LS-SVM is a very promising machine learning technique and will gain more extensive application.

References

1. Boobis, A., Gundert-Remy, U., Kremers, P., Macheras, P. and Pelkonen, O., Eur. J. Pharm. Sci., 17 (2002) 183.
2. Baláž, Š. and LukaÂcÏova, V., Quant. Struct. Act. Relat., 18 (1999) 361.
3. Zhang, H.B., J. Pharm. Sci., 93 (2004) 1595.
4. Zhang, H.B., J. Chem. Inf. Comput. Sci., 45 (2005) 121.
5. Ekins, S. and Obach, R.S., J. Pharmacol. Exp. Ther., 295 (2000) 463.
6. Cronin, M.T.D., Curr. Opin. Drug Discovery Dev., 3 (2000) 292.
7. Yoshida, F. and Topliss, J.G., J. Med. Chem., 43 (2000) 2575.
8. Belousov, A.I., Verzakov, S.A. and Von Frese, J., Chemometr. Intell. Lab. Syst., 64 (2002) 15.
9. Morris Colin, W., Autret, A. and Boddy, L., Ecol. Model., 146 (2001) 57.
10. Burbidge, R., Trotter, M., Buxton, B. and Holden, S., Comput. Chem., 26 (2001) 5.
11. Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Chem. Inf. Comput. Sci., 43 (2003) 1288.
12. Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Chem. Inf. Comput. Sci., 44 (2004) 161.
13. Xue, C.X., Zhang, R.S., Liu, H.X., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Chem. Inf. Comput. Sci., 44 (2004) 669.
14. Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Comput. Aid. Mol. Des., 18 (2004) 389.
15. Yao, X.J., Panaye, A., Doucet, J.P., Zhang, R.S., Chen, H.F., Fan, B.T., Liu, M.C. and Hu, Z.D., J. Chem. Inf. Comput. Sci., 44 (2004) 1257.
16. Liu, H.X., Xue, C.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Chem. Inf. Comput. Sci., 44 (2004) 1979.
17. Burges, C.J.C., Data Min. Know. Disc, 2 (1998) 1.
18. Vapnik, V. Estimation of Dependences Based on Empirical Data. Springer, Berlin, 1982.
19. Suykens, J.A.K. and Vandewalle, J., Neural Process. Lett, 9 (1999) 293.
20. Katritzky, A.R., Lobanov, V.S. and Karelson, M., Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual. Version 2.0, 1994.
21. Katritzky, A.R., Lobanov, V.S. and Karelson, M., Chem. Soc. Rev., 24 (1995) 279.
22. Oblak, M., Randic, M. and Solmajer, T., J. Chem. Inf. Comput. Sci., 40 (2000) 994.
23. Katritzky, A.R. and Tatham, D.B., J. Chem. Inf. Comput. Sci., 41 (2001) 1162.
24. HyperChem. 4.0, Hypercube, 1994.
25. Stewart, J.P.P. MOPAC 6.0, Quantum Chemistry Program Exchange; QCPE, No. 455. Indiana University, Bloomington, IN, 1989.
26. Katritzky, A.R., Petrukhin, R., Jain, R. and Karelson, M., J. Chem. Inf. Comput. Sci., 41 (2001) 1521.
27. Cortes, C. and Vapnik, V., Machine Learn., 20 (1995) 273.
28. Vapnik, V. Statistical Learning Theory. Wiley, New York, 1998.
29. Schölkopf, B., Burges, C. and Smola, A. Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA, 1999.
30. Cristianini, N. and Shawe-Taylor, J. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000.
31. URL: http://www.kernel-machines.org/,Dec, 2004.
32. Pelckmans, K., Suykens, J.A.K., Van Gestel, T., De Brabanter, D., Lukas, L., Hamers, B., De Moor, B. and Vandewalle, J., LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines. Internal Report 02-44, ESATSISTA; K.U. Leuven, Leuven, 2002.
33. Stanton, D.T., Egolf, L.M. and Jurs, P.C.J., Chem. Inf. Comput. Sci., 32 (1992) 306.