Journal of Molecular Graphics and Modelling 29 (2010) 214–220

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Prediction of blood–brain partitioning: A model based on molecular electronegativity distance vector descriptors

Yong-Hong Zhang [a,b,c], Zhi-Ning Xia [a,**], Li-Tang Qin [c], Shu-Shen Liu [c,*]

[a] College of Bioengineering, Chongqing University, Chongqing 400030, People's Republic of China
[b] College of Pharmaceutical Sciences, Chongqing Medical University, Chongqing 400016, People's Republic of China
[c] State Key Laboratory of Pollution Control and Resources Reuse, Key Laboratory of Yangtze River Water Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, Shanghai 200092, People's Republic of China

ARTICLE INFO

Article history:
Received 7 September 2009
Received in revised form 14 June 2010
Accepted 17 June 2010
Available online 25 June 2010

Keywords:
QSAR
Molecular electronegativity distance vector (MEDV)
PLSR
The variable importance in projection (VIP)
Blood–brain barrier (BBB) permeability

ABSTRACT

The objective of this paper is to build a reliable model based on the molecular electronegativity distance vector (MEDV) descriptors for predicting the blood–brain barrier (BBB) permeability and to reveal the effects of the molecular structural segments on the BBB permeability. Using 70 structurally diverse compounds, the partial least squares regression (PLSR) models between the BBB permeability and the MEDV descriptors were developed and validated by the variable selection and modeling based on prediction (VSMP) technique. The estimation ability, stability, and predictive power of a model are evaluated by the estimated correlation coefficient (r), leave-one-out (LOO) cross-validation correlation coefficient (q), and predictive correlation coefficient (R_p). It has been found that PLSR model has good quality, $r = 0.9202$, $q = 0.7956$, and $R_p = 0.6649$ for M1 model based on the training set of 57 samples. To search the most important structural factors affecting the BBB permeability of compounds, we performed the values of the variable importance in projection (VIP) analysis for MEDV descriptors. It was found that some structural fragments in compounds, such as $-CH_3$, $-CH_2-$, $=CH-$, $=C=$, $\equiv C-$, $-CH<$, $=C<$, $=N-$, $-NH-$, $=O$, and $-OH$, are the most important factors affecting the BBB permeability.

1. Introduction

Accurate prediction of pharmacokinetic properties, such as absorption, distribution, metabolism, excretion, and toxicity (ADMET), plays a very important role in very early stage of drug discovery to reduce the failure rate of drug candidates in clinical trials [1,2]. Blood–brain barrier (BBB) is a complex cellular system consisting of endothelial cells of the brain capillaries, which separates the brain and central nervous system (CNS) from the bloodstream [3]. BBB permeability is one of the most important pharmacokinetic properties for a drug candidate, especially for the CNS drugs. This BBB barrier must be crossed for the drugs aimed at CNS targets, while the non-CNS drugs should be retained in the bloodstream to minimize the undesired CNS side effects. Usually the blood–brain partition coefficient log BB ($BB = C_{brain}/C_{blood}$) is used to determine whether a compound could cross the BBB or not, where C_{brain} and C_{blood} are the steady-state/equilibrium concentrations of the compound in brain and blood, respectively [4]. Experimental determination of log BB of a compound, however, is quite complicated and difficult.

Quantitative structure-activity/property relationship (QSAR/QSPR) models offer such *in silico* screening by predicting log BB from the molecular structure of the compounds, hence supporting the "fail fast, fail cheap" business model of drug development [5]. Levin [6] and Young et al. [7] found a good correlation between log BB and log P (octanol/water partition coefficient). Abraham et al. developed descriptors based on hydrogen-bond and polar surface area (PSA) (such as hydrogen-bond acidity, solute dipolarity/polarizability, hydrogen-bond, and log P) to relate to the brain–blood partition [8–11]. Norinder et al. [12] used the program MolSurf to compute theoretical molecular descriptors related to the physico-chemical properties, such as lipophilicity, polarity, polarizability, and hydrogen-bonding, and developed a partial least square regression (PLSR) model predicting the log BB of organic solutes. Clark [13] reported the correlation of log BB with PSA and log P. Luco [14] used several topological and constitutional descriptors to model the brain–blood concentration ratio. Many other researchers also made contributions to the model establishments summarized in the review [15,16]. Chen et al. [17] found that the PSA seems to be the most important factor for BBB permeability.

* Corresponding author. Tel.: +86 021 65982767.
** Co-corresponding author. Tel.: +86 023 65106615.
E-mail addresses: zhnxia@yahoo.com.cn (Z.-N. Xia), ssliuhl@263.net (S.-S. Liu).

1093-3263/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.jmgm.2010.06.006

However, the development of a reliable predictive model requires more rational and more precise descriptors to reflect molecular properties. Obviously, the descriptors related to physicochemical properties are limited in the application in the theoretical models. In recent years, the theoretical descriptors based on 2D or 3D molecular structures have been rapidly developed and a lot of practicable software/methods which are used to compute the descriptors and/or model QSAR/QSPR had been come forth, such as DRAGON [18], TSAR [19], and CoMFA/CoMSIA [20–22]. Ooms et al. developed a simple model to predict log BB from 3D molecular fields [23]. Katritzky et al computed structural descriptors based on molecule and on fragment using CODESSA-PRO and ISIDA programs to give QSAR models [24]. Obrezanova et al. employed the Gaussian process regression to develop automatic QSAR for the evaluation of ADMET properties such as the log BB [25,26]. The common modeling methods that researches used were multiple linear regression (MLR) [27–30], PLSR [31,32], *k*-nearest neighbors (*k*NN) [5,33], and support vector machine [33,34].

It is still necessary to develop a simple model with the structure-based computational descriptors for predicting the log BB of the structurally diverse drugs. It is well known that the biological activity of a molecule depends on the structure such as the atomic electronegativity and the distance between the atoms. In this paper, we chose the molecular electronegativity distance vector (MEDV) descriptors [35–37] to characterize a set of the structurally diverse compounds. The significance of the MEDV descriptors is their consideration of the electrical characteristics of each atom and the distance between atoms in organic chemicals. The MEDV descriptors were directly and rapidly computed from two-dimensional topological structure of a molecule and applied in the QSAR/QSPR studies on many complicated molecular systems, such as cyclooxygenase-2 (COX-2) inhibitors [38], polychlorinated biphenyls (PCBs) [39–40], polybrominated diphenyl ethers (PBDEs) [41], non-ionic organic compounds (NOC) [42], and non-polar organic compounds (NPOC) [43]. In this paper, we employed the PLSR to address the collinearity or auto-correlation problems among the MEDV descriptors and use the variable selection and modeling based on prediction (VSMP) technique [44] developed in our laboratory to select the number of the latent variables in PLSR analysis. Then, a PLSR model between the log BB and MEDV descriptors was developed and validated. The estimation ability, stability, and predictive power of the model developed are evaluated by the estimated correlation coefficient (r), leave-one-out (LOO) cross-validation correlation coefficient (q), and predictive correlation coefficient (R_p). Furthermore, we analyzed the most important structural factors affecting the log BB of a compound according to the values of the variable importance in projection (VIP) [14,45].

2. Materials and methods

In this study, the development of models consists of three stages: (a) entry and storage of molecular structures as well as generation of MEDV descriptors derived directly from the molecular topological structures, (b) the optimization of the latent variables in PLSR models by VSMP, and (c) the development and validation of a prediction model between the log BB and some MEDV descriptors.

2.1. Data set

70 compounds under study and their log BB values were directly taken from the literature [13]. The serial numbers and the experimental log BB values of the compounds were listed in Table 1. The log BB values are widespread and distributed in the range of −2.15 to 1.04. These compounds have diverse structures including many

Table 1
The log BB values of 70 compounds where OBS refers to log BB observed, M1 to log BB calculated by the M1 model, and the samples with an asterisk (*) to those in the test set.

No	Compound in the literature [13]	log BB values		
		OBS	M1	M1-OBS
1*	1	−1.42	−0.79	0.63
2	2	−0.04	−0.31	−0.27
3	3	−2.00	−1.88	0.12
4	4	−1.30	−1.30	0.00
5	5	−1.06	−1.30	−0.24
6	6	0.11	−0.04	−0.15
7*	7	0.49	0.94	0.45
8	8	0.83	1.11	0.28
9	9	−1.23	−0.88	0.35
10	10	−0.82	−1.20	−0.38
11	11	−1.17	−0.56	0.61
12*	12	−2.15	−0.72	1.43
13	13	−0.67	−0.52	0.15
14	14	−0.66	−0.55	0.11
15	15	−0.12	−0.32	−0.20
16	16	−0.18	−0.64	−0.46
17	17	−1.15	−0.85	0.30
18	18	−1.57	−1.22	0.35
19	19	−1.54	−1.36	0.18
20*	20	−1.12	−0.64	0.48
21	21	−0.73	−0.44	0.29
22	22	−0.27	−0.29	−0.02
23	23	−0.28	−0.42	−0.14
24	24	−0.46	−0.47	−0.01
25	25	−0.24	−0.40	−0.16
26	26	−0.02	−0.09	−0.07
27*	27	0.69	−0.08	−0.77
28*	28	0.44	−0.36	−0.80
29	29	0.14	−0.41	−0.55
30*	30	0.22	0.60	0.38
31	Butanone	−0.08	−0.18	−0.10
32	Benzene	0.37	0.16	−0.21
33	3-Methylpentane	1.01	0.68	−0.33
34*	3-Methylhexane	0.90	0.73	−0.17
35	2-Propanol	−0.15	−0.21	−0.06
36	2-Methylpropanol	−0.17	0.06	0.23
37	2-Methylpentane	0.97	0.63	−0.34
38	2,2-Dimethylbutane	1.04	1.22	0.18
39	1,1,1-Trifluoro-2-chloroethane	0.08	0.21	0.13
40	1,1,1-Trichloroethane	0.40	0.07	−0.33
41	Diethyl ether	0.00	−0.17	−0.17
42	Enflurane	0.24	0.44	0.20
43	Ethanol	−0.16	−0.02	0.14
44	Fluroxene	0.13	0.22	0.09
45*	Halothane	0.35	0.29	−0.06
46	Heptane	0.81	0.56	−0.25
47	Hexane	0.80	0.53	−0.27
48	Isoflurane	0.42	0.45	0.03
49	Methane	0.04	0.07	0.03
50*	Methylcyclopentane	0.93	0.29	−0.64
51	Nitrogen	0.03	−0.05	−0.08
52	Pentane	0.76	0.50	−0.26
53	Propanol	−0.16	0.07	0.23
54	Propanone	−0.15	−0.30	−0.15
55	Teflurane	0.27	0.28	0.01
56	Toluene	0.37	0.37	0.00
57	Trichloroethene	0.34	0.42	0.08
58	Y-G14	−0.30	−0.19	0.11
59*	Y-G15	−0.06	0.20	0.26
60	Y-G16	−0.42	−0.28	0.14
61	Y-G19	−1.30	−0.39	0.91
62*	Y-G20	−1.40	0.12	1.52
63	SKF89124	−0.43	0.04	0.47
64	SKF101468	−0.25	0.20	0.45
65	31	0.00	−0.47	−0.47
66	32	−0.34	−0.37	−0.03
67	33	−0.30	−0.62	−0.32
68	34	−1.34	−1.16	0.18
69	35	−1.82	−2.11	−0.29
70*	36	0.87	−0.03	−0.90

drug categories, such as adrenergic, analgesic, antiviral, barbiturate, histamine, neuroleptic, benzodiazepine, etc. It should be noted that the log BB value of compound 70 is 0.87, which is an average value from the literature [13] (0.76–0.98).

2.2. Structure coding and MEDV descriptors calculation

The molecular structures of all compounds were entered into the computer by input of atomic attributes of all non-hydrogen atoms in the molecules and adjacency relationships between all pairs of atoms as structure code. 91 MEDV descriptors denoting the interaction between two atomic types were then calculated according to MEDV theory [36,37,43]. 43 atomic attributes and 13 atomic types for various non-hydrogen atom in organic molecules were defined in our previous papers [36,38,43]. The calculation procedure of the MEDV descriptors can be simply stated as follows. Firstly, the intrinsic state (I) of a non-hydrogen atom is computed using Eq. (1) [36,43]:

$$I = \sqrt{\frac{\nu}{4}}\frac{(2/n)^2\delta^\nu + 1}{\delta} = \sqrt{\frac{\nu}{4}}\frac{(2/n)^2(\sigma+\pi-h)+1}{\sigma-h} \quad (1)$$

where ν is the number of valence electrons; n the principal quantum number for the valence shell of that atom; σ and π are, respectively, the number of electrons in σ and π orbitals; h is the number of hydrogen atoms bonded to the atom. Secondly, the relative electronegativity (Q) of the non-hydrogen atom is calculated using Eq. (2) [36,43]:

$$Q_i = I_i + \sum_{j\neq i}^{\text{all } j}\frac{(I_i - I_j)}{d_{ij}^2} \quad (2)$$

where d_{ij} is the shortest graph distance between atom i and j. Finally, the MEDV descriptor, x_z, is calculated using Eq. (3) [36,43]:

$$x_z = m_{kl} = \sum_{i\in k, j\in l}\frac{Q_iQ_j}{d_{ij}^2} \quad (k, l=1, 2, \ldots, 13;\ \ l=k;\ \ z=1, 2, \ldots, 91) \quad (3)$$

where k and l are the atomic types of the ith and jth non-hydrogen atoms, which i and j are the serial number of the non-hydrogen atoms in a molecule, respectively; z is the serial number of the MEDV descriptors.

2.3. Variable selection and modeling

For a real compound system, the value of one or many MEDV descriptors could be zero for all molecule samples due to absence of one or many atomic types and the descriptors should be deleted before modeling. Furthermore, if one descriptor only has a few non-zero samples such as 1, 2, 3, 4, and 5, the descriptor has statistically little meaning and should be deleted prior to developing a model. After above deletion, a PLSR model was developed by VSMP technique [44] developed in our laboratory. The developing and validation process of log BB models in this paper were shown in Fig. 1.

Firstly, 57 samples were randomly selected from the data set consists of 70 compounds as a training set and the remaining 13 ones as a test set. For the training set of 57 samples, the MEDV descriptors only with 0, 1, 2, 3, 4, or 5 non-zero samples were deleted from the pool of 91 MEDV descriptors. Using the VSMP technique and taking the q obtained in the LOO cross-validation as a criterion, the number of the latent variables (A) in a PLSR model was selected. An optimal PLSR model was then built and validated by using the test set of 13 samples. The randomization selection, descriptor deletion, and model development above were repeated 10 times.

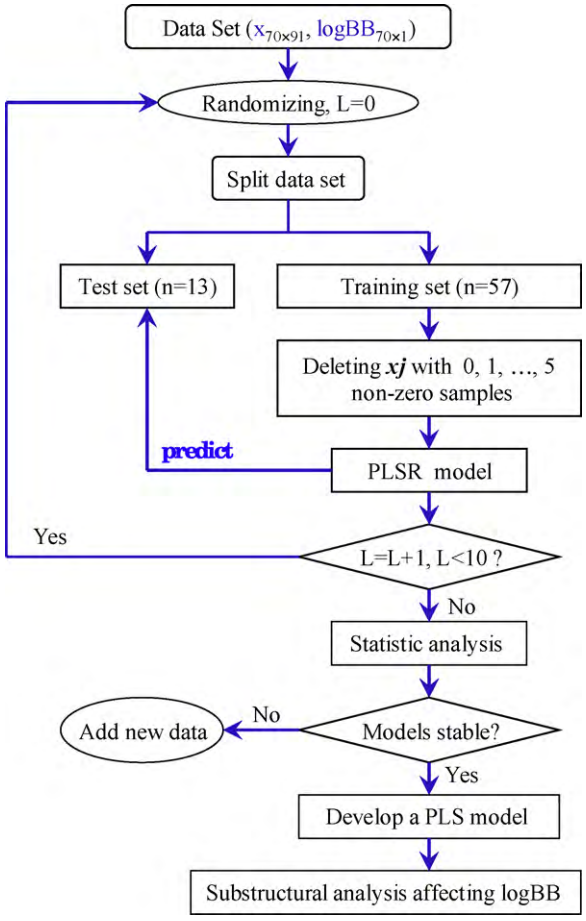

Fig. 1. Sketch map for modeling and validation process of log BB.

Secondly, the statistical analysis on 10 PLSR modeling was performed. The average values and 95% confidence intervals of r and *RMSE* were respectively calculated to observe the stability of the PLSR models with MEDV descriptors. If the PLSR models are relatively stable, a PLSR model was randomly chosen to turn to next step. Using the VSMP technique, the PLSR analysis on a training set was performed and the effects of the important MEDV descriptors or molecular substructures on the log BB values were rationally discussed.Thirdly, the absolute errors (AEs) between the calculated log BB values and experimental ones of compounds in the test set were calculated. If all AEs of a compound in all test sets are larger than $3 \times RMSE$, the compound will affect the stability and predictive potential of the model and its structure should be analyzed.

3. Results and discussion

3.1. Statistic analysis of PLSR models

To develop and validate a credible QSAR model, the data set consisting of 70 compounds having 91 MEDV descriptors was randomly split into a training set of 57 samples and a test set of 13 samples and repeated 10 times. After deleting the variables with a few non-zero samples having little significance statistically, the VSMP technique was used to select A and to develop PLSR model on the training set of 57 samples having m non-zero descriptors. Table 2 listed some statistics, such as m (the number of the MEDV descriptors in the training set), A (the number of the latent variables), r (the estimated correlation coefficient), and *RMSE* (the estimated root-mean-square error), and the test set compounds in 10 modeling.

Table 2
Some statistics, such as *m*, *A*, *r*, and *RMSE*, and the test set compounds in 10 modeling.

Model	*m*	*A*	*r*	*RMSE*	Compounds in test set
M1	39	4	0.9202	0.28	34,30,*70*,59,*12*,7,20,27,*62*,45,1,50,28
M2	36	4	0.9134	0.31	66,36,30,28,*67*,45,26,3,13,*68*,48,*69*,*70*
M3	37	4	0.9195	0.28	*10*,*69*,1,*30*,50,*26*,37,2,38,22,60,3,*62*
M4	38	4	0.9118	0.31	51,24,63,*62*,50,39,*10*,46,3,66,68,32,25
M5	38	4	0.9322	0.27	58,59,55,6,50,43,23,16,*69*,2,*62*,*12*,*61*
M6	37	4	0.9206	0.27	19,34,52,20,63,*12*,29,*68*,40,*67*,32,*69*,11
M7	39	4	0.9143	0.30	56,26,33,*12*,69,59,63,*11*,43,*7*,70,25,*61*
M8	38	4	0.9153	0.32	*62*,45,17,21,31,64,41,16,60,39,28,*12*,26
M9	38	4	0.9353	0.27	47,28,58,42,68,53,50,*11*,63,60,*12*,*30*,*62*
M10	40	4	0.9331	0.26	3,36,9,60,68,*70*,63,26,*11*,32,*10*,18,*62*
Mean			0.9216	0.287	

From Table 2, the number of non-zero descriptors in the training set ranges from 36 to 40, which shows the structural diversity of compounds selected randomly. However, the number of the latent components obtained in 10 PLSR analysis is $A = 4$. The average values of *r* and *RMSE* were 0.9216 (±0.0063) and 0.287 (±0.022) where the numerical value in parenthesis refers to 95% confidence interval, respectively, which explain good estimation ability and stability of 10 PLSR models (called M1 to M10) built randomly. Clearly, the statistical qualities of the PLSR models are close to or better than those obtained in previous studies (Table 4). It was reasonable to use the MEDV descriptors to estimate the BBB permeability of compounds.

From Table 2, it could be found that a few compounds (italic) entered into the test set have higher AEs than $3 \times RMSE$. The compound 12 is randomly selected six times in 10 modeling and six AEs are larger than $3 \times RMSE$. Likewise, the AEs of compound 62 randomly selected seven times in 10 modeling are all larger than $3 \times RMSE$. It is indicated that two compounds (nos. 12 and 62) could affect the stability of the model and predictive potential. However, the AEs of other compounds selected many times in 10 random test sets, such as compound 10, 11, 26,. . ., and 70, are not all larger than $3 \times RMSE$, which implies that the compounds do not always affect the quality of model.

3.2. An example of PLSR model

For the convenience of use, we chose the first PLSR model (noted as M1) in Table 2 as an example to give more detailed discussion. In the M1 model development, the variety of the *r* and *RMSE* in modeling and the *q* and *RMSV* in the LOO validation with *A* was listed in Table 3. It has shown *A* of 4. Some statistics obtained in modeling, validation, and prediction were given as follows:

$n = 57$, $A = 4$, $m = 39$, $r = 0.9202$, $RMSE = 0.28$, $F = 71.87$ (modeling)
$n = 57$, $A = 4$, $m = 39$, $q = 0.7956$, $RMSV = 0.44$ (LOO cross-validation)
$n_p = 13$, $A = 4$, $m = 39$, $R_p = 0.6649$, $RMSP = 0.78$ (test set prediction)

where *n* and n_p are the number of compounds in the training set and test set, and *F* and *RMSP* are the Fischer's statistic and the predictive root-mean-square error, respectively.

Model M1 has a good estimation ability ($r = 0.9202$, $RMSE = 0.28$), high stability ($q = 0.7956$, $RMSV = 0.44$) and predictive potential ($R_p = 0.6649$, $RMSP = 0.78$). The log BB values calculated by M1 (column M1) were listed in Table 1 together with the log BB observed (column OBS). It should be indicated why the R_p is lower and *RMSP* is higher is because the predictive error for the log BB of compound 12 is 1.43 and one of compound 62 is 1.52 from Table 1.

The plot of log BB values calculated by M1 vs. those observed was shown in Fig. 2. In Fig. 2, two points are far away from the diagonal line, such as the compounds of nos. 12 and 62 in the test set. Except for the compounds of nos. 12 and 62, others data points are almost symmetrically distributed around the diagonal line, which indicated that PLSR model based on 39 MEDV descriptors has a good estimation ability and has predict potential for the BBB permeability of compounds.

The comparison of quality of our model M1 with some linear models [26,33,46] studied previously was shown in Table 4. For the same data set as reported by Clark [13], the estimation ability of M1 are better than the best model DEC-II ($r^2 = 0.7868$, $RMSE = 0.35$) in Clark's work. The quality of our model M1 is close to or better than that of many linear models listed in Table 4. The 8-descriptor model reported by Ma et al. [30] had very high fitting capability ($n = 37$, $r^2 = 0.9120$, $RMSE = 0.23$), but their model did not run suitable validation and the ratio of molecules to descriptors (often called the Topliss ratio [47]) was 4.625 which is less than the ratio of at least 5 required in one normal model. It should be indicated that the non-linear *k*NN-SVM made by Zhang et al. [33] was excellent and included the most samples ($n = 144$) in training set. However, there are too descriptors such as 324, 184, and 346 ones in their models. Al-Fahemi et al. [48] employed the momentum-space descriptors to predict BBB permeability and developed the 12-descriptor model with a very high fitting capability ($n = 42$, $r^2 = 0.943$, $RMSE = 0.16$ and $F = 40$), but they did not gave the stability and predictive potential of the model and the Topliss ratio was too small (3.5).

Table 3
The variety of some statistics such as *r*, *RMSE*, *q*, and *RMSV* with *A* for the training set of 57 compounds (model M1 in Table 2 where $n = 57$ and $m = 39$).

A	*r*	*RMSE*	*q*	*RMSV*
1	0.8047	0.43	0.7481	0.48
2	0.8629	0.36	0.7736	0.45
3	0.8991	0.31	0.7859	0.45
4	0.9202	0.28	0.7956	0.44
5	0.9380	0.25	0.7940	0.45
6	0.9493	0.23	0.7798	0.49

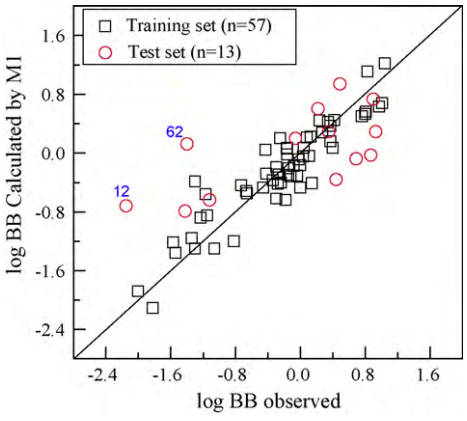

Fig. 2. Plot of the log BB calculated by the model M1 vs. the log BB observed.

Table 4
QSAR model comparison evaluating log BB values of compounds where *m* and *n* denote the number of descriptors and the number of samples used in modeling.

Resources	Group	*m*	*A*	Descriptors	*n*	r^2	q^2	*RMSE*	n_p	*RMSP*
	Young	1		log *P*	20	0.69		0.427		
	Kansy	2		PSA and molecular volume	20	0.697		0.437		
	Abraham	5		Molecular property descriptors	57		0.91	0.195		
	Lombardo	1		Free energy of solvation	55	0.67		0.406	6	
	Clark	2		PSA and log *P*	55	0.787		0.351	5	
	Luco	25	18	Topological and constitutional descriptors	58	0.850	0.752	0.315	22	0.408
	Feher	3		Number of hydrogen-bond acceptors, log *P*, PSA	61	0.73	0.688	0.424	25	0.789
Editing from Table	Kelder	1		Dynamic PSA	45	0.841				
in literature [33]	Norinder	14	3	Molecular property descriptors	28	0.862	0.782	0.29	28	0.353
	Platts	6		Molecular property descriptors	148	0.745	0.711	0.341		
	Keseru	1		Solvation free energies	55	0.72		0.367	5	
	Salminen	3		Lipophilicity, molecular size and acid/base character	23	0.848		0.313		
	Ma	8		Inter- and intra-molecular solute descriptors	37	0.912		0.229	8	
	Katritzky	5		*c* log *P*, etc.	113	0.781	0.752	0.35	19	0.18
	Hou	3		High-charged PSA, log *P*, MW360	72	0.785		0.355		
	Iyer	5		PSA, *c* log *P* and membrane-solute descriptors	56	0.845	0.795		7	
	Pan	2		TPSA and *c* log *P*	37	0.85	0.83		46	
	Subramanian	8		*A* log *P*	58	0.845	0.811	0.311	39	0.463
	Rose	3		E-state index and molecular connectivity index	102	0.66	0.62	0.478	28	
	Winkler	7		Property-based descriptors	106	0.81	0.65	0.37		
Zhang [33]	Zhang	324		MolConnZ 4.05 descriptors, MOE descriptors, Dragon descriptors	144	0.91			15	
Narayanan [46]	Narayanan	4		Kappa shape index, E-state index, topological descriptor, log *P*	88	0.7461	0.7177	0.392		
Obrezanova [26]	Obrezanova	7		SMARTS based descriptors	106	0.79		0.32	22	0.49
This paper	M1	39	4	MEDV descriptors	57	0.8468	0.6329	0.28	13	0.78

3.3. Structure analysis

The MEDV descriptors, shown in our recent report [43], could be represented as an interaction between two atomic types. The structures of the compounds under study are very diverse due to too many atomic types, which require many MEDV descriptors to distinguish between the structures of the compounds. The VIP values of the descriptors in PLSR analysis were used to explain the importance of the descriptors contributing to log BB [14]. The higher the VIP value of a variable is, the higher its contribution to the model is. The descriptor with a small VIP value does not contain much useful information for modeling the BBB permeability and cannot be considered. One or many descriptors with the minimum VIP were in turn deleted from the training set to examine the variety of the statistic quality (such as q). If there is no difference between the model quality especially the LOO validation statistics before and after deletion, the deletion can be carried out. Once the difference begins become significant the deletion stops. In this way, 17 important MEDV descriptors whose VIP values were greater than 1.000 were extracted from the set of 39 descriptors in model M1. Column diagram of the values of VIP for 17 variables (VIP > 1.000) were showed in Fig. 3.

The most important structural factors affecting the log BB of a compound were discussed according to the VIP values of 17 MEDV descriptors. From Fig. 3, though there were 17 variables whose VIP values are bigger than 1.000, five of the VIP values are bigger than 1.500 and much higher than the others obviously. According to the MEDV theory [37,43], the five important MEDV descriptors in the M1 come from the interactions between pairs of atomic types, nos. 6 and 6 (x_{56}), nos. 6 and 9 (x_{59}), nos. 3 and 6 (x_{29}), nos. 2 and 6 (x_{18}), and nos. 1 and 6 (x_6), respectively, i.e., the 5 descriptors relate to only 5 atomic types of nos. 1, 2, 3, 6, and 9, especially for atomic type of nos. 6. Every atomic type relates to one atom segment or atom group. For 70 compounds under study, five atomic types correspond to five atom segments, 1 (–CH_3), 2 (–CH_2–, =CH–, =C= and ≡C–), 3 (–CH< and =C<), 6 (=N–, –NH–), and 9 (=O, –OH), respectively. To explain the effect of each atomic type on the BBB permeability, the most important descriptor is the 56th MEDV descriptor implying an interaction between the atomic type 6 (=N–, –NH–) and 6 (=N–, –NH–). When all five important descriptors are emerged in one compound, the predictive error is less (i.e. compound 3 and 4). Both compound 12 and 62 have only one atom segment of =N– ($x_{56} = 0$), so their predictive errors are much larger. As an example, the molecular structure, atomic types, and MEDV descriptors of the compound 3 were illustrated in Fig. 4.

The results above showed that the atom segments, –CH_3, –CH_2–, =CH–, =C=, ≡C–, –CH<, =C<, =N–, –NH–, =O, and –OH, are the most important factors affecting the BBB permeability, especially for =N– and –NH–. Our MEDV descriptors can intuitively analyze the relationship between molecular structure and the BBB permeability.

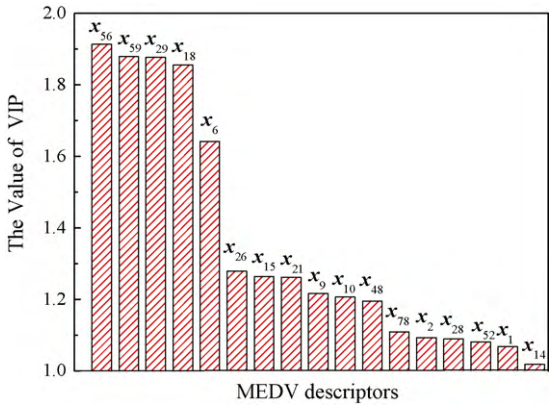

Fig. 3. Histogram of the values of VIP for 17 variables (VIP > 1.00) calculated by the model M1.

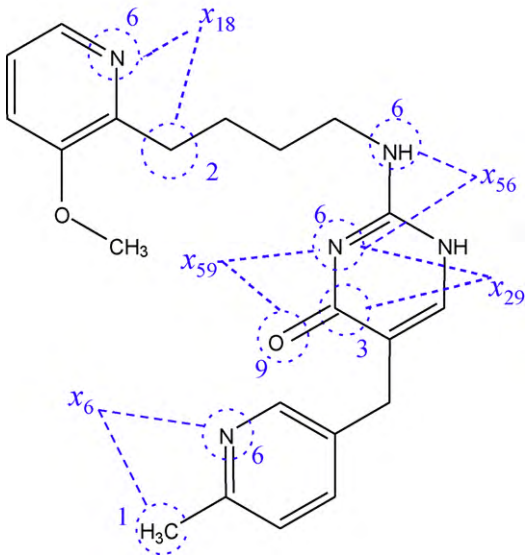

Fig. 4. The molecular structure, atomic types, and MEDV descriptors of the compound of no. 3 in Table 1 (Arabic numeral refers to atomic type of a non-hydrogen atom; x_6, x_{18}, x_{29}, x_{56}, and x_{59} refer to MEDV descriptors representing the interaction between two atomic types).

However, our MEDV descriptors are still two-dimensional structural variables and could not in detail interpret the mechanism of BBB permeability in biophysical and medicinal chemical.

4. Conclusions

The PLSR models evaluating the BBB permeability were derived and validated by using the MEDV descriptors of 70 compounds. The PLSR models (M1) have a good estimation ability, high stability, and proper predictive power. Some important descriptors affecting BBB permeability were picked up by the VIP ordering analysis. It has been found that main structural factors influencing the BBB permeability of compounds are the molecular substructures, $-CH_3$, $-CH_2-$, =CH–, =C=, ≡C–, –CH<, =C<, =N–, –NH–, =O and –OH.

Acknowledgments

We are especially grateful to the Major National Science and Technology Project of P.R. China (2008ZX07421-001) and the Foundation for the Author of National Excellent Doctoral Dissertation of P.R. China (200355) and the Key Laboratory of Yangtze River Water Environment Foundation (YRWEY1002) for their financial supports.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2010.06.006.

References

[1] I.V. Tetko, P. Bruneau, H.W. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME-Tox predictions? Drug Discov. Today 11 (2006) 700–707.
[2] R.F. Liu, H.M. Sun, S.S. So, Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood–brain barrier penetration, J. Chem. Inf. Comput. Sci. 41 (2001) 1623–1632.
[3] P.L. Golden, G.M. Pollack, Blood–brain barrier efflux transport, J. Pharm. Sci. 92 (2003) 1739–1753.
[4] J. Shen, Y.P. Du, Y.X. Zhao, G.X. Liu, Y. Tang, In silico prediction of blood–brain partitioning using a chemometric method called genetic algorithm based variable selection, QSAR Comb. Sci. 27 (2008) 704–717.
[5] D.A. Konovalov, D. Coomans, E. Deconinck, Y. Vander Heyden, Benchmarking of QSAR models for blood–brain barrier permeation, J. Chem. Inf. Model. 47 (2007) 1648–1656.
[6] V.A. Levin, Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability, J. Med. Chem. 23 (1980) 682–684.
[7] R.C. Young, R.C. Mitchell, T.H. Brown, C.R. Ganellin, R. Griffiths, M. Jones, K.K. Rana, D. Saunders, I.R. Smith, Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H_2 receptor histamine antagonists, J. Med. Chem. 31 (1988) 656–671.
[8] M.H. Abraham, Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical process, Chem. Soc. Rev. 22 (1993) 73–83.
[9] M.H. Abraham, H.S. Chadha, R.C. Mitchell, Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain, J. Pharm. Sci. 83 (1994) 1257–1268.
[10] M.H. Abraham, H.S. Chadha, R.C. Mitchell, Hydrogen-bonding. Part 36. Determination of blood brain distribution using octanol–water partition coefficients, Drug Des. Discov. 13 (1995) 123–131.
[11] M.H. Abraham, K. Takacs-Novak, R.C. Mitchell, On the partitioning of ampholytes: application to blood–brain distribution, J. Pharm. Sci. 86 (1997) 310–315.
[12] U. Norinder, P. Sjoberg, T. Osterberg, Theoretical calculation and prediction of brain–blood partitioning of organic solutes using MolSurf parametrization and PLS statistics, J. Pharm. Sci. 87 (1998) 952–959.
[13] D.E. Clark, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration, J. Pharm. Sci. 88 (1999) 815–821.
[14] J.M. Luco, Prediction of the brain–blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling, J. Chem. Inf. Comput. Sci. 39 (1999) 396–404.
[15] U. Norinder, M. Haeberlein, Computational approaches to the prediction of the blood–brain distribution, Adv. Drug Deliv. Rev. 54 (2002) 291–313.
[16] J.T. Goodwin, D.E. Clark, In silico predictions of blood–brain barrier penetration: considerations to "Keep in mind", J. Pharmacol. Exp. Ther. 315 (2005) 477–483.
[17] Y. Chen, Q.J. Zhu, J. Pan, Y. Yang, X.P. Wu, A prediction model for blood–brain barrier permeation and analysis on its parameter biologically, Comput. Methods Prog. Biomed. 95 (2009) 280–287.
[18] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON Software for the Calculation of Molecular Descriptors ver. 5.2 for Windows, Talete S.r.l., Milan, Italy, 2005.
[19] Accelrys, TSAR 3.3 Software, Oxford Molecular Limited, Oxford, England, 2000.
[20] W.J. Geldenhuys, P.R. Lockman, T.H. Nguyen, C.J. Van der Schyf, P.A. Crooks, L.P. Dwoskin, D.D. Allen, 3D-QSAR study of bis-azaaromatic quaternary ammonium analogs at the blood–brain barrier choline transporter, Bioorg. Med. Chem. 13 (2005) 4253–4261.
[21] W.J. Geldenhuys, P.R. Lockman, J.H. McAfee, K.T. Fitzpatrick, C.J. Van der Schyf, D.D. Allen, Molecular modeling studies on the active binding site of the blood–brain barrier choline transporter, Bioorg. Med. Chem. Lett. 14 (2004) 3085–3092.
[22] I. Lessigiarska, I. Pajeva, M.T. Cronin, A.P. Worth, 3D QSAR investigation of the blood–brain barrier penetration of chemical compounds, SAR QSAR Environ. Res. 16 (2005) 79–91.
[23] F. Ooms, P. Weber, P.A. Carrupt, B. Testa, A simple model to predict blood–brain barrier permeation from 3D molecular fields, BBA-Mol. Basis Dis. 1587 (2002) 118–125.
[24] A.R. Katritzky, M. Kuanar, S. Slavov, D.A. Dobchev, D.C. Fara, M. Karelson, W.E. Acree, V.P. Solov'ev, A. Varnek, Correlation of blood–brain penetration using structural descriptors, Bioorg. Med. Chem. 14 (2006) 4888–4917.
[25] O. Obrezanova, G. Csanyi, J.M.R. Gola, M.D. Segall, Gaussian processes: a method for automatic QSAR modeling of ADME properties, J. Chem. Inf. Model. 47 (2007) 1847–1857.
[26] O. Obrezanova, J.M.R. Gola, E.J. Champness, M.D. Segall, Automatic QSAR modeling of ADME properties: blood–brain barrier penetration and aqueous solubility, J. Comput. Aid Mol. Des. 22 (2008) 431–440.
[27] M.C. Hutter, Prediction of blood–brain barrier permeation using quantum chemically derived information, J. Comput. Aid Mol. Des. 17 (2003) 415–443.
[28] S. Van Damme, W. Langenaeker, P. Bultinck, Prediction of blood–brain partitioning: a model based on ab initio calculated quantum chemical descriptors, J. Mol. Graph. Model. 26 (2008) 1223–1236.
[29] X.C. Fu, Z.F. Song, C.Y. Fu, W.Q. Liang, A simple predictive model for blood–brain barrier penetration, Pharmazie 60 (2005) 354–358.
[30] X.L. Ma, C. Chen, J. Yang, Predictive model of blood–brain barrier penetration of organic compounds, Acta Pharmacol. Sin. 26 (2005) 500–512.
[31] M. Feher, E. Sourial, J.M. Schmidt, A simple model for the prediction of blood–brain partitioning, Int. J. Pharm. 201 (2000) 239–247.
[32] P. Crivori, G. Cruciani, P.A. Carrupt, B. Testa, Predicting blood–brain barrier permeation from three-dimensional molecular structure, J. Med. Chem. 43 (2000) 2204–2216.
[33] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, QSAR Modeling of the blood–brain barrier permeability for diverse organic compounds, Pharm. Res. 25 (2008) 1902–1914.
[34] U. Norinder, Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection, Neurocomputing 55 (2003) 337–346.
[35] S.S. Liu, C.Z. Cao, Z.L. Li, Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, λ, J. Chem. Inf. Comput. Sci. 38 (1998) 387–394.

[36] S.S. Liu, C.S. Yin, Z.L. Li, S.X. Cao, QSAR study of steroid benchmark and dipeptides based on MEDV-13, J. Chem. Inf. Comput. Sci. 41 (2001) 321–329.
[37] S.S. Liu, C.S. Yin, L.S. Wang, Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors, J. Chem. Inf. Comput. Sci. 42 (2002) 749–756.
[38] S.S. Liu, S.H. Cui, D.Q. Yin, Y.Y. Shi, L.S. Wang, QSAR studies on the COX-2 inhibition by 3,4-diarylcycloxazolones based on MEDV descriptor, Chin. J. Chem. 21 (2003) 1510–1516.
[39] L.T. Qin, S.S. Liu, H.L. Liu, H.L. Ge, A new predictive model for the bioconcentration factors of polychlorinated biphenyls (PCBs) based on the molecular electronegativity distance vector (MEDV), Chemosphere 70 (2008) 1577–1587.
[40] S.S. Liu, Y. Liu, D.Q. Yin, X.D. Wang, L.S. Wang, Prediction of chromatographic relative retention time of polychlorinated biphenyls from the molecular electronegativity distance vector, J. Sep. Sci. 29 (2006) 296–301.
[41] Y.H. Zhang, S.S. Liu, H.Y. Liu, Predicting the gas chromatographic relative retention time of polybrominated diphenyl ethers by MEDV-13 descriptors, Chromatographia 65 (2007) 319–324.
[42] S.S. Liu, L.T. Qin, H.L. Liu, D.Q. Yin, Molecular electronegativity distance vector model for the prediction of bioconcentration factors in fish, J. Mol. Model. 14 (2008) 83–92.
[43] L.T. Qin, S.S. Liu, H.L. Liu, QSPR model for bioconcentration factors of nonpolar organic compounds using molecular electronegativity distance vector descriptors, Mol. Divers. 14 (2010) 67–80.
[44] S.S. Liu, H.L. Liu, C.S. Yin, L.S. Wang, VSMP: a novel variable selection and modeling method based on the prediction, J. Chem. Inf. Comput. Sci. 43 (2003) 964–969.
[45] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[46] R. Narayanan, S.B. Gunturi, In silico ADME modelling: prediction models for blood–brain barrier permeation using a systematic variable selection method, Bioorgan. Med. Chem. 13 (2005) 3017–3028.
[47] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure–activity relationships, J. Med. Chem. 22 (1979) 1238–1244.
[48] J.H.A. Al-Fahemi, D.L. Cooper, N.L. Allan, Investigating the utility of momentum-space descriptors for predicting blood–brain barrier penetration, J. Mol. Graph. Model. 26 (2007) 607–612.