Journal of Computer-Aided Molecular Design **19:** 33–46, 2005.
DOI 10.1007/s10822-005-0095-8

The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine

H.X. Liu[1], R.J. Hu[1], R.S. Zhang[1,2], X.J. Yao[1,3], M.C. Liu[1], Z.D. Hu[1,*] & B.T. Fan[3]
[1]*Department of Chemistry, Lanzhou University, Lanzhou 730000, P.R. China;* [2]*Department of Computer Science, Lanzhou University, Lanzhou 730000, P.R. China;* [3]*Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, F-75005 Paris, France*

Received 25 August 2004; accepted in revised form 3 January 2005
© Springer 2005

Key words: heuristic method, human oral absorption, QSPR/QSAR, support vector machine

Summary

Support vector machine (SVM), as a novel machine learning technique, was used for the prediction of the human oral absorption for a large and diverse data set using the five descriptors calculated from the molecular structure alone. The molecular descriptors were selected by heuristic method (HM) implemented in CODESSA. At the same time, in order to show the influence of different molecular descriptors on absorption and to well understand the absorption mechanism, HM was used to build several multivariable linear models using different numbers of molecular descriptors. Both the linear and non-linear model can give satisfactory prediction results: the square of correlation coefficient R^2 was 0.78 and 0.86 for the training set, and 0.70 and 0.73 for the test set respectively. In addition, this paper provides a new and effective method for predicting the absorption of the drugs from their structures and gives some insight into structural features related to the absorption of the drugs.

Introduction

Successful drug development requires not only optimization of specific and potent pharmacological activity, but also efficient drug delivery to the target site. Many drug candidates fail to reach their therapeutic potentials due to poor biolovailability [1]. To improve the quality of ''research'' compounds, efficient methods are required in early drug discovery to search key factors influencing drug absorption and understand the absorption mechanisms responsible for drug transport, which are of significant pharmaceutical importance. At the same time, with the growth of combinatorial chemistry methods in drug discovery, a large number of candidate compounds are synthesized and screened in parallel for *in vitro* pharmacological activity, which has dramatically increased the demand for rapid and efficient models for estimating human intestinal absorption [2]. As a consequence, the prediction of human intestinal absorption has been a major goal in the design, optimization, and selection of candidates for the development of oral drugs. However, the investigations of drug oral absorption in humans and animals are time-consuming and somewhat difficult to perform. Although cell membrane methods and *in vivo* animal studies have been used instead of human intestinal absorption methods, these techniques are still costly and labor intensive [2].

Alternatively, quantitative structure–property/activity relationship (QSPR/QSAR) provides a promising method for the estimation of absorption behavior of drugs based on the descriptors derived solely from the molecular structure to fit experimental data. The advantage of this approach over other methods lies in the fact that it requires only

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

the knowledge of chemical structure and is not dependent on the experiment properties. This approach is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with changes in molecular features of the compounds termed descriptors [3]. 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. It is generally assumed that physicochemical descriptors of drug molecules can be useful for predicting absorption for passive diffusion of drugs. Consequently, QSPR/QSAR have been successfully established to predict human intestinal absorption [2, 4–12]. However, some of models are only applicable to the limited range and can not be applied to a larger data set [4, 5, 8, 12]. Some models are only established using the simple linear and non-linear correlations based on the fairly simple molecular descriptors such as $\log P$, polar surface area, H-bonding acceptors and donors, and Abraham descriptors [2, 7, 11]. However, generally, the factors influencing absorption behavior of drug were complex and not all of them were linear correlated with the absorption. Thus, it is difficult to obtain the accurate predictive results by using these models. Due to the above reason, it is very necessary to build the accurate quantitative models to predict the absorption of drug based on a diverse dataset by using the powerful and robust QSPR/QSAR techniques.

Among the investigation of QSPR/QSAR, one of the important factors affecting the quality of the 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. Various numerical representations of the compounds were proposed in QSPR/QSAR studies, such as: constitutional and topological descriptors; numerical code; quantum chemistry descriptors, etc. The Software CODESSA, developed by the 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 [13, 14]. 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 [15, 16].

Another important factor responsible for the quality of the QSPR/QSAR model is the method to build the model. In response to increased accuracy demands, artificial intelligence techniques have been applied to QSPR/QSAR analysis since the late 1980s [17]. Machine learning techniques have, in general, offered greater accuracy than have their statistical forebears, but there exist accompanying problems for the QSPR/QSAR analyst to consider. Neural networks, for example, offer high accuracy in most cases but can suffer from the reproducibility of results, due largely to random initialization of the network and variation of stopping criteria, and lack of information regarding the classification produced [18]. Genetic algorithms can suffer in a similar manner. The stochastic nature of both population initialization and the genetic operators used during training can make results hard to reproduce [19]. Owing to the reasons outlined above; there is a continuing need for the application of more accurate and informative techniques to QSPR/QSAR analysis.

The support vector machine (SVM) is a popular algorithm developed from the machine learning community at present. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application [17, 20–28].

In the present investigation, SVM, as a novel machine learning technique, for the first time, was used for the prediction of the human oral absorption based on the large and diverse data set using the descriptors calculated from the molecular structure alone by the software CODESSA. Five descriptors were selected as inputs by heuristic method (HM). In order to investigate the

influence of different descriptors on absorption and to well understand the absorption mechanism, HM was used to build several multivariable linear models. The aim of this study was to explore the absorption behavior of the drugs with diverse structures and establish a new and accurate quantitative structure–absorption relationship model and to confirm the possibility of predicting drug absorption and to seek for the structural factors affecting their absorption. The prediction results are satisfactory in both training set and test set compounds, which proved SVM was a useful tool in the prediction of the drug absorption.

Methods

Data set

The human intestinal absorption dosed orally of 169 drugs was collected from [7] and is listed in Table 1. The data set was randomly separated into a training set of 113 compounds and a test set of 56 compounds. The training set was used to build the model, and the test set was used to evaluate its prediction ability. Leave-one-out (LOO) cross-validation was performed for the whole training set.

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. A more precise optimization was done with semi-empirical AM1 method in MOPAC. 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.) [13].

The heuristic method [13]

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 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 cross-validated R^2_{cv}, and the F-value).

The heuristic method usually produces correlations 2–5 times faster than other methods with comparable quality [29]. The rapidity of calculations from the heuristic method renders it the first method of choice in practical research. Thus, in

Table 1. Predicted absorption by HM and SVM.

No.	Name	% Abs.	Results of SVM		Results of MLR	
			Predicted	Residue	Predicted	Residue
Training set						
1	Aminopyrine	100	98	−2	95	−5
2	Caffeine	100	89	−11	77	−23
3	Camazepam	100	95	−5	99	−1
4	Cisapride	100	87	−13	87	−13
5	Corticosterone	100	101	1	86	−14
6	Desipramine	100	98	−2	105	5
7	Diazepam	100	99	−1	101	1
8	Ethinylestradiol	100	97	−3	94	−6
9	Fenclofenac	100	96	−4	95	−5
10	Gallopamil	100	96	−4	99	−1
11	Glyburide	100	81	−19	80	−20
12	Imipramine	100	102	2	102	2
13	Indomethacin	100	95	−5	93	−7
14	Ievonorgestrel	100	99	−1	111	11
15	Iormetazepam	100	94	−6	98	−2
16	Mexiletine	100	100	−0	96	−4
17	Nefazodone	100	99	−1	93	−7
18	Ondansetron	100	101	1	104	4
19	Oxatomide	100	101	1	98	−2
20	Piroxicam	100	92	−8	90	−10
21	Praziquantel	100	97	−3	91	−9
22	Salicylicacid	100	90	−10	80	−20
23	Stavudine	100	99	−1	80	−20
24	Tenoxicam	100	90	−10	88	−12
25	Testosterone	100	101	1	107	7
26	Toremifene	100	105	5	110	10
27	Valproicacid	100	94	−6	93	−7
28	Carfecillin	99	86	−13	76	−23
29	Naproxen	99	97	−2	96	−3
30	Prenisolone	99	83	−16	76	−23
31	Propranolol	99	93	−6	93	−6
32	Lamotrigine	98	97	−1	82	−16
33	Minoxidilne	98	97	−1	87	−11
34	Viloxazine	98	94	−4	94	−4
35	Warfarin	98	97	−1	94	−4
36	Clafibrate	97	96	−1	101	4
37	Disulfiram	97	96	−1	87	−10
38	Venlafaxine	97	95	−2	96	−1
39	Bumetanide	96	75	−21	72	−24
40	Trapidil	96	96	−0	91	−5
41	Codeine	95	94	−1	92	−3
42	Flumazenil	95	93	−2	90	−5
43	Ibuprofen	95	98	3	107	12
44	Metaprolol	85	90	5	90	5
45	Oxprenolol	95	89	−6	88	−7
46	Scopolamine	95	94	−1	96	1
47	Sotalol	95	94	−1	93	−2

Table 1. (Continued).

No.	Name	% Abs.	Results of SVM		Results of MLR	
			Predicted	Residue	Predicted	Residue
48	Alprenolol	93	92	−1	90	−3
49	Amrinone	93	85	−8	80	−13
50	Ketoprofen	92	97	5	103	11
51	Kydiocortisone	91	90	−1	78	−13
52	Alprazolam	90	105	15	106	16
53	Anphentamine	90	97	7	99	9
54	Chloramphenicol	90	79	−11	71	−19
55	Felbamate	90	84	−6	90	0
56	Meloxicam	90	91	1	100	10
57	Misoldipine	90	95	5	93	3
58	Phenytoin	90	92	2	97	7
59	Sulindac	90	100	10	98	8
60	Tramadol	90	96	6	92	3
61	Sihydrocodeine	89	94	5	82	−7
62	Sultopride	89	90	1	85	−4
63	Tenidap	89	88	−1	76	−12
64	Mitrendipine	88	84	−4	100	12
65	Nitrendipine	88	95	7	89	2
66	Bupropion	87	95	8	57	−30
67	Lamivudine	87	65	−22	71	−15
68	Topiramate	86	78	−8	94	9
69	Lansoprazole	85	92	7	91	6
70	Oxyfedrine	85	94	9	82	−3
71	Tolbutamide	85	84	−1	95	11
72	Brumazepam	84	97	13	83	−1
73	Captopril	84	85	1	75	−7
74	Methylprednisolone	82	80	−2	82	−0
75	Mifobate	82	81	−1	60	−21
76	Digoxin	81	80	−1	72	−9
77	Flecainkde	81	80	−1	96	15
78	Quinidine	81	97	16	84	4
79	Acebutolol	80	81	1	75	−5
80	Dexamethasone	80	80	0	66	−14
81	Ethambutol	80	74	−6	82	2
82	Isoniazid	80	91	11	103	23
83	Methadone	80	100	20	77	−1
84	Urapidil	78	84	6	81	4
85	Famciclovir	77	82	5	77	1
86	Propylthiouracil	76	79	3	69	−4
87	Cycloserine	73	74	1	67	−2
88	Hydrochlorothiazide	69	70	1	72	8
89	Cimetidine	64	81	17	73	11
90	Terburaline	62	74	12	72	11
91	Furosemide	61	72	11	63	3
92	Pirbuterol	60	68	8	62	2
93	Reproterol	60	61	1	74	17
94	Nadolol	57	73	16	91	34
95	Sumatriptan	57	93	36	49	−1

Table 1. (Continued).

No.	Name	% Abs.	Results of SVM		Results of MLR	
			Predicted	Residue	Predicted	Residue
96	Amiloride	50	49	−1	74	24
97	Atenolol	50	76	26	69	21
98	Rimiterol	48	70	22	71	24
99	Cymarin	47	73	26	69	25
100	Sulpiride	44	78	34	50	12
101	Famotidine	38	39	1	34	3
102	Fosfomycin	31	30	−1	27	−3
103	Fosmisomycin	30	27	−3	55	27
104	Metivudine	28	51	23	19	2
105	Foscarnet	17	18	1	27	11
106	K-strophanthoide	16	17	1	53	50
107	Mannitol	16	15	−1	−10	−12
108	Ganciclovir	3	47	44	3	2
109	Acarbose	2	1	−1	−2	−3
110	Kanamycin	1	1	0	23	23
111	Neomycin	1	2	1	0	0
112	Lactulose	0.6	17	16	81	−9
113	Raffinose	0.3	−2	−2	31	15
Test set						
114	Bornaprine	100	101	1	101	1
115	Cicaprost	100	91	−9	87	−13
116	Cyproterone	100	98	−2	103	3
117	Diclofenac	100	96	−4	100	0
118	Fluvastatin	100	86	−14	84	−16
119	Granisetron	100	99	−1	96	−4
120	Isoxicam	100	91	−9	91	−9
121	Lornoxicam	100	92	−8	91	−9
122	Nicotine	100	98	−2	92	−8
123	Phenglutarimide	100	90	−10	78	−22
124	Progesterone	100	113	13	102	2
125	Sudoxicam	100	92	−8	90	−10
126	Theophylline	100	86	−14	72	−28
127	Verapamil	100	97	−3	100	0
128	Nordiazepam	99	98	−1	97	−2
129	Atropine	98	95	−3	97	−1
130	Tolmesxide	98	97	−1	99	1
131	Antipyrine	97	98	1	99	2
132	Trimethoprim	97	86	−11	80	−17
133	Torasemide	96	81	−15	79	−17
134	Fluconazole	95	94	−1	98	3
135	Labetalol	95	73	−22	69	−26
136	Practolol	95	78	−17	76	−19
137	Timolol	95	78	−17	74	−21
138	Isradipine	92	94	2	99	7
139	Naloxone	91	86	−5	83	−8
140	Betaxolol	90	91	1	92	2
141	Ketorolac	90	101	11	105	15
142	Nizatidine	90	95	5	85	−5

Table 1. (Continued).

No.	Name	% Abs.	Results of SVM		Results of MLR	
			Predicted	Residue	Predicted	Residue
143	Terazosin	90	83	−7	80	−10
144	Oxazepam	89	93	4	94	5
145	Felodipine	88	93	5	98	10
146	Saccharin	88	94	6	84	−4
147	Pindolol	87	87	−0	85	−2
148	Morphine	85	88	3	85	−0
149	Acetylsalicylicacid	84	95	11	93	9
150	Propiverine	84	101	17	97	13
151	Sorivydine	82	46	−36	52	−30
152	Piroximone	81	91	10	78	−3
153	Zcetaminophen	80	91	11	85	5
154	Gauanmbenz	80	85	5	78	−2
155	Omeprazole	80	95	15	96	16
156	Mercaptoethamesulfonicacid	77	86	9	79	2
157	Recainam	71	86	15	83	12
158	Metolazone	64	82	18	81	17
159	Fenoterol	60	68	8	69	9
160	Ziprasidone	60	97	37	89	29
161	Metformin	53	59	6	53	−0
162	Guanoxan	50	80	30	76	26
163	Metaproterenol	44	74	30	73	29
164	Ascorbic	35	56	21	58	23
165	Lincomycin	28	55	27	51	23
166	Adefovir	16	23	7	38	22
167	Cidofovir	3	11	8	27	24
168	Ouabain	1.4	35	33	38	37
169	Streptomycin	1	6	5	−2	−3

the present investigation, we used this method to build the linear model.

Support vector machine

Since the factors influencing absorption behavior of drug were complex and not all of them were linear correlation with the absorption, in order to build the more accuracy predictive model, it is necessary to build non-linear model. SVM, developed by Vapnik [30], as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. Comparing with traditional neural networks, SVM possesses prominent advantages: (1) strong theoretical background provides SVM with high generalization capability and can avoid local minima; (2) SVM always has a solution, which can be quickly obtained by a standard algorithm (quadratic programming); (3) SVM need not determine network topology in advance, which can be automatically obtained when the training process ends; (4) SVM builds a result based on a sparse subset of training samples, which reduces the workload. Originally, SVM are developed for pattern recognition problems, such as image recognition [31], microarray gene expression classification [17], protein folding recognition [32], protein structural class prediction [33], identification of protein cleavage sites, QSAR and other pharmaceutical data analysis [17, 34]. And now, with the introduction of insensitive loss function, SVM have been extended to solve non-linear regression estimation and time-series prediction with excellent performances [35]. Based on the reason outlined above, support vector

regression was used to build the non-linear model, whose basic principle was described as belows.

There exist a number of excellent introductions into SVM [36–38]. For this reason, we will only briefly describe the main ideas of SVM regression here.

A support vector machine is first trained on a sample with objects having known target values. After training, the machine is used to predict or estimate target values for objects where these values are unknown. A kernel-induced feature space with function $k(x_i, x)$ is used for the mapping of objects onto target values. Thus a non-linear feature mapping will allow the treatment of non-linear problems in a linear space. The prediction or approximation function used by a basic SVM is

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \qquad (1)$$

where α_i is some real value, $\mathbf{x}_i$ is a feature vector corresponding to a training object, and $k(x_i, x)$ is a kernel function. The components of vector α and the constant b represent the hypothesis and are optimized during training. $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function, which value is equal to the inner product of two vectors $\mathbf{x}$ and $\mathbf{x}_i$ in the feature space $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}_i)$. That is, $K(\mathbf{x}, \mathbf{x}_i) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i)$. 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 $\Phi(\mathbf{x})$ explicitly and it may be useful to think of the kernel, $K(\mathbf{x}, \mathbf{x}_i)$ as comparing patterns, or as evaluating the proximity of objects in their feature space. Thus a test point is evaluated by comparing it to all training points. Training points with non-zero weight α_i are called the *support vectors*.

For a given dataset, only the kernel function and the regularity parameter C must be selected to specify one SVM. Any function that satisfies Mercer's condition can be used as the kernel function. In support vector regression, the Gaussian kernel $K(u, v) = \exp(-|u - v|^2/\delta^2)$ is most commonly used.

SVM implementation and computation environment

All calculation programs implementing SVM were written in R-file based on R script for SVM. All scripts were compiled using R1.7.1 compiler running operating system on a Pentium IV with 256 M RAM.

Results and discussion

The heuristic method model

A total of 728 descriptors were calculated by the CODESSA program for all the compounds. After the heuristic reduction, the pool of descriptors was reduced to 230. To select the set of descriptors that are most relevant to the absorption of drugs and show the affecting degree for absorption of different descriptors and well understand of the absorption mechanism, the linear models with 1–8 variables were built. The best 1–8 parameter models are listed as follows:

1-Parameter model

$$\%\ \mathrm{ABS} = 1.5775 \times 10^2 - 3.6367 \times 10^3 \mathrm{FPSA3}$$

$$R^2 = 0.6513, R^2_{\mathrm{cv}} = 0.5900, F = 211.17,$$

$$s^2 = 253.3162$$

2-Parameter model

$$\%\ \mathrm{ABS} = 1.1769 \times 10^2 - 2.6285 \times 10^3 \mathrm{HACA2/TMSA} - 2.6767 \times 10^1 \mathrm{FPSA}$$

$$R^2 = 0.6929, R^2_{\mathrm{cv}} = 0.6556, F = 124.08,$$

$$s^2 = 238.0091$$

3-Parameter model

$$\%\ \mathrm{ABS} = 1.0569 \times 10^2 - 2.1624 \times 10^3 \mathrm{HACA2/TMSA} - 5.5142 \times 10^1 \mathrm{FPSA2} + 3.2287 \mathrm{RI3}$$

$$R^2 = 0.7465, R^2_{\mathrm{cv}} = 0.6921, F = 107.02,$$

$$s^2 = 198.2151$$

4-Parameter model

$$\%\ \mathrm{ABS} = 9.5465 \times 10^1 - 1.4772 \times 10^3 \mathrm{HACA2/TMSA} - 1.9831 \times 10\mathrm{TEI} + 1.1278 \times 10^{-1} \mathrm{TMEI} + 4.2515 \times 10^2 \mathrm{MNRIC}$$

$$R^2 = 0.7647, R^2_{\mathrm{cv}} = 0.7195, F = 87.74,$$

$$s^2 = 185.7357$$

5-Parameter model

$$\%\ \mathrm{ABS} = 9.2049 \times 10^{1} - 1.5516 \times 10^{3}\mathrm{HACA2/TMSA} - 1.8968 \times 10\mathrm{TEI} + 1.1490 \times 10^{-1}\mathrm{TMEI} + 3.9425 \times 10^{2}\mathrm{MNRIC} + 3.9191 \times 10\mathrm{RP}$$

$$R^2 = 0.7811, R^2_{\mathrm{cv}} = 0.7345, F = 76.36, s^2 = 174.4017$$

6-Parameter model

$$\%\ \mathrm{ABS} = 9.7496 \times 10^{1} - 2.3185 \times 10^{3}\mathrm{HACA2/TMSA} - 5.0387 \times 10\mathrm{FPSA2} + 2.7690\mathrm{RI3} + 8.2256 \times 10^{2}\mathrm{MNRIC} - 2.6067 \times 10^{3}\mathrm{ANRIC} + 2.2457 \times 10^{2}\mathrm{MIA}$$

$$R^2 = 0.7930, R^2_{\mathrm{cv}} = 0.7241, F = 67.67, s^2 = 166.4841$$

7-Parameter model

$$\%\ \mathrm{ABS} = 3.3383 \times 10^{2} - 1.9991 \times 10^{3}\mathrm{HACA2/TMSA} - 1.6278 \times 10\mathrm{TEI} + 7.6854 \times 10^{-2}\mathrm{TMEI} + 7.9150 \times 10^{2}\mathrm{MNRIC} + 3.6247 \times 10\mathrm{RP} - 2.7133 \times 10^{3}\mathrm{ANRIC} - 2.9612 \times 10^{1}\mathrm{MASEH}$$

$$R^2 = 0.8055, R^2_{\mathrm{cv}} = 0.7431, F = 62.14, s^2 = 157.8695$$

8-Parameter model

$$\%\ \mathrm{ABS} = 3.6564 \times 10^{2} - 2.0426 \times 10^{3}\mathrm{HACA2/TMSA} - 1.5481 \times 10\mathrm{TEI} + 6.5454 \times 10^{-2}\mathrm{TMEI} + 8.4731 \times 10^{2}\mathrm{MNRIC} + 3.2612 \times 10\mathrm{RP} - 3.0217 \times 10^{3}\mathrm{ANRIC} - 2.9612 \times 10^{1}\mathrm{MASEH} - 4.1333 \times 10\mathrm{IOKSE}$$

$$R^2 = 0.8142, R^2_{\mathrm{cv}} = 0.7470, F = 56.95, s^2 = 152.3205$$

The involved molecular descriptors and their corresponding physical–chemical meaning are given in Table 2. Figure 1 shows the plots of R^2 and R^2_{cv} for the training set as a function of the number of descriptors for the 1–8-parameter models. R^2 increased with increasing the number of descriptors. However, the values of R^2_{cv} decreased for the 6-parameter model, which suggested the overfitting of data happened with the 6–8-parameter models. The 5-parameter model was chosen as the best linear model and the corresponding descriptors were used as inputs for the non-linear model. In the following, we only discussed 1–5-parameter models in detail.

By interpreting the descriptors in the regression models, it is possible to gain some insight into factors that are likely to govern the absorption of

Table 2. The involved molecular descriptors and their corresponding physical–chemical meaning.

Symbol	Physical–chemical meaning
FPSA3	FPSA-3 fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]
HACA2/TMSA	HACA-2/TMSA [quantum-chemical PC]
FPSA2	FPSA-2 fractional PPSA (PPSA-2/TMSA) [Zefirov's PC]
RI3	Randic index (order 3)
TEI	Topographic electronic index (all bonds) [Zefirov's PC]
TMEI	Tot. molecular electrostatic interaction
MNRIC	Max. nucleoph. react. index for a C atom
RP	Polarity parameter/square distance
ANRIC	Avg nucleoph. react. index for a C atom
MIA	Moment of inertia A
MASEH	Max. atomic state energy for a H atom
IOKSE	Image of the Onsager–Kirkwood solvation energy

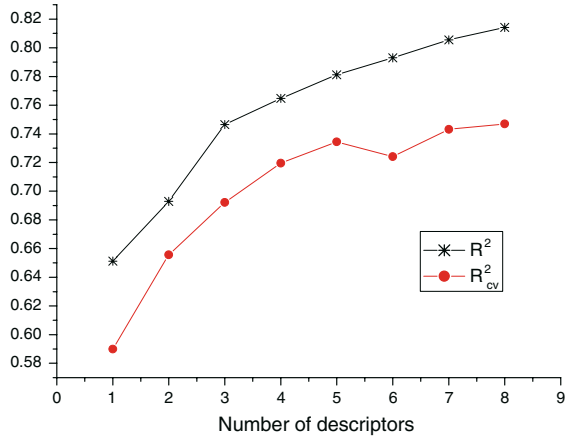

Figure 1. Influence of the number of descriptors on the correlation coefficient (R^2) and the cross-validation correlation coefficient (R^2_{cv}) of the regression models.

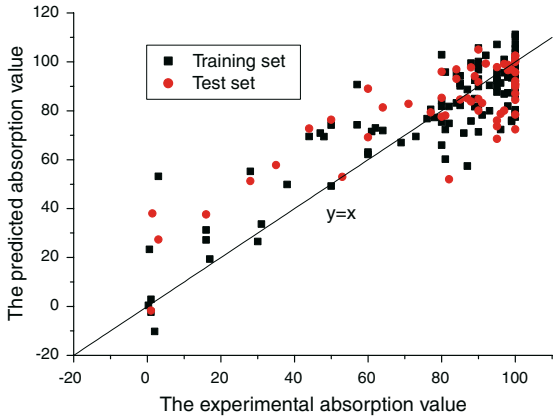

Figure 2. Plot of predicted absorption values vs. experimental values for the training set and test set based on the 5-parameter model by heuristic method.

the drugs and understand which interaction play an important role during the absorption process of the series of drugs.

Oral absorption refers to the movement of a drug from its site of administration into the blood. Sietsema [39] defined absorption as "the drug passing from the lumen of the gastrointestinal (GI) tract into the tissue of the GI tract. Once in the tissue, the drug is considered to be absorbed". The steps involved in the absorption of an orally administered drug can be simply depicted as follows:

$$\text{Solid drug} \overset{\text{dissolution}}{\longrightarrow} \text{Drug in solution}$$

$$\overset{\text{Diffusion}}{\underset{\text{Through GI}}{\longrightarrow}} \text{Absorbed drug} \overset{\text{Perfusion}}{\underset{\text{In portal vein}}{\longrightarrow}}$$

According to the difference of rate-limited step, the drugs can be classified as dissolution rate-limited drugs, diffusion rate-limited drugs and perfusion rate-limited drugs. In this investigation, all 169 drugs belong to diffusion rate-limited drugs [7]. For the diffusion rate-limited drugs, the rate of absorption is mainly determined by the permeability characteristics of the drug. Thus, all the molecular descriptors which can influence the permeability characteristics of the drug will influence their absorption.

From the 1-parameter model, we can see the most important molecular descriptor is the fractional partial positive surface area, FPSA-3 fractional PPSA. FPSA3 [40] is defined as the ratio of the atomic charge weighted partial positive surface area (PPSA3), which is obtained by summation of products of the individual atomic partial charges and the atomic solvent-accessible surface areas, and the total molecular surface area (TMSA). The high correlation between this descriptor and the absorption demonstrates that the absorption of drug is largely dependent on accessible surface area and the localized partial charges of the molecules. In addition, this descriptor can be loosely related to the hydrogen bonding acceptor and donor ability, indicating that the hydrogen bond also plays an important role during the absorption of drugs. The negative regression coefficient of this descriptor indicates an increase in its value leads to a decrease in the value of absorption.

For the 2-parameter model, two CPSA descriptors FPSA2 and HACA-2/TMSA are involved. FPSA2 [16], FPSA-2 Fractional PPSA (PPSA-2/TMSA), which is defined as molecule weighted fractional positive charged surface area, has the similar meaning to the FPSA3. The correlation coefficient between the two descriptors is up to 0.895. HACA-2/TMSA Hydrogen bond Acceptor Charged surface Area/Total Molecular Surface Area (HACA-1/TMSA), describes the hydrogen bonding acceptor properties of the compounds. The great improvement of the quality of model proved further the importance of the hydrogen bond.

Compared with the 2-parameter model, Randic index 3 (RI3), was added in the 3-parameter model. The Randic index [41] is calculated as a

sum of atomic connectivities over molecular paths of certain length $(1, 2, \ldots, n)$, it thus reflects molecular size and branching and then reflects the solubility of the drug. The introduction of this descriptor results in more than 0.05 unit improvement of R^2 and 40 unit decrease of s^2, which showed the solubility also played an important role during the absorption. The positive regression coefficient indicates that the larger the value of this descriptor, the stronger the solubility of the drug and the stronger the absorption, which is in agreement with [42].

In the 4-parameter model, four descriptors belong to electrostatic and quantum chemical descriptors, respectively. The electrostatic descriptor Topographic Electronic Index (TEI) [41], a complex function of the atomic charges and molecular geometry, is calculated as the sum of the charge differences over all molecular bonds. The quantum chemical descriptor Total Molecular Electrostatic Interaction [13] (TMEI), characterizes the total energy of the molecule in electrostatic energy scales and describes the electrostatic feature of the molecule. These two descriptors can account for the polar interaction during the absorption. Another quantum chemical descriptor Max Nucleoph. React. Index (MNRIC) for a C atom [13], is a reactivity indice descriptor which estimates the relative reactivity of the atoms in the molecule for a given series of compounds and is related to the activation energy of the corresponding chemical reaction. Since most atoms are the C atoms in the present investigation, MNRIC can be responsible for the reactivity of compounds. The positive correlation coefficients showed the absorption increased as the value of this descriptor increased. It can be explained as below: the larger the value, the more easily the compound can interact with body fluid and biology membrane and then the better the absorption.

On the basis of the 4-parameter, the 5-parameter model more roundly considered the polar interaction by the introduction of the relative polarity parameter (RP) Polarity parameter/square distance, and the quality of model becomes better.

From the above discussion, we concluded that the factors influencing the absorption of the series of drugs mainly included the polar, electrostatic, hydrogen bonds, solubility features of drugs and the selected descriptors were able to account for these features. The best 5-parameter model gave an root mean square (rms) error of 12.85 for the training set, 14.96 for the prediction set and the corresponding correlation coefficients (R^2) were 0.78, 0.70, respectively, confirming the predictive capability of the model. The predicted values for all the 169 compounds of training set and test set were given in Table 1. Figure 2 showed the plot of the calculated vs. experimental absorption for all the 169 compounds studied, the training set and the test set.

Result of SVM

Since the factors influencing the absorption of these compounds could be complex and not all of them were linear correlation with the absorption, SVM was used to build the non-linear predictive model to further discuss the correlation between the molecular structure and the absorption based on the selected descriptors of the 5-parameter model.

SVM parameters optimization

Similar with other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter C, ε of ε-insensitive loss function, the kernel type K and its corresponding parameters. C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small then insufficient stress will be placed on fitting the training data. If C is too large then the algorithm will overfit the training data. But, [38] indicated that prediction error was scarcely influenced by C. In order to make the learning process stable, a large value should be set up for C (e.g., $C = 100$).

The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of ε is critical from theory.

The kernel type is another important one. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function in R is as follows:

$$\exp(-\gamma^*|u - v|^2)$$

where γ is a constant, the parameter of the kernel; u, v are two independent variables; γ controls the amplitude of the Gaussian function and therefore, controls the generalization ability of SVM. We have to optimize γ and find the optimal one.

In order to find the optimum values of two parameters (γ and ε) and prohibit the overfitting of the model, the leave-one-out cross-validation of the training set was performed. The LOO procedure consists of removing one example from the training set, constructing the decision function on the basis only of the remaining training data and then testing on the removed example. In this fashion one tests all examples of the training data and measures the fraction of errors over the total number of training examples. The square of correlation coefficient base on the LOO cross-validation method (R^2_{cv}) was used as an evaluating function.

Detailed process of selecting the parameters and the effects of every parameter on generalization performance of the corresponding model are shown in Figures 3 and 4. To obtain the optimal γ, the support vector learning machines with different γ were trained. We calculated the R^2_{cv} on different γ, according to the generalization ability of the model based on the LOO cross-validation for the training set in order to determine the optimal one. The curve of R^2_{cv} vs. gamma is shown in Figure 3. The optimal γ was found as 0.021. In order to find an optimal ε, the R^2_{cv} on different ε was calculated. The curve of the R^2_{cv} vs. the epsilon is shown in Figure 4. From Figure 4, the optimal ε was found as 0.15.

The predicted result of SVM

From the above discussion, the values of γ, ε and C were fixed to 0.021, 0.15 and 100, respectively. The corresponding number of support vectors was 80. The predicted results of the optimal SVM are shown in Table 1 and Figure 5. The model gave an rms of 10.35 for the training set, 14.08 for the prediction set and the corresponding correlation coefficients (R^2) were 0.86, 0.73, respectively. From Table 1 and Figure 5, it can be seen that the predicted values are in agreement with the experimental values for most of drugs.

Compared with the results from the HM, the rms errors of the SVM model for the training, the test set were lower than that of the HM. The correlation coefficient (R^2) given by SVM model were higher than that of the HM.

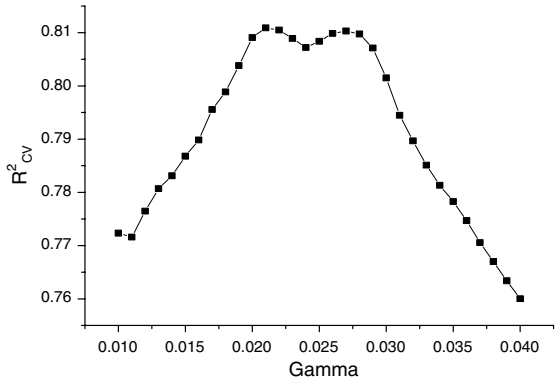

Figure 3. R^2_{cv} vs. gamma (C=100, ε = 0.05).

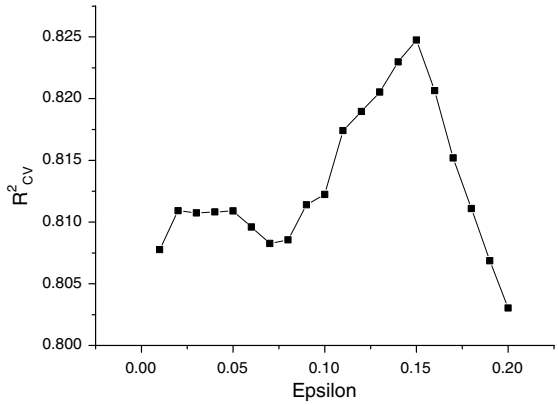

Figure 4. R^2_{cv} vs. epsilon.

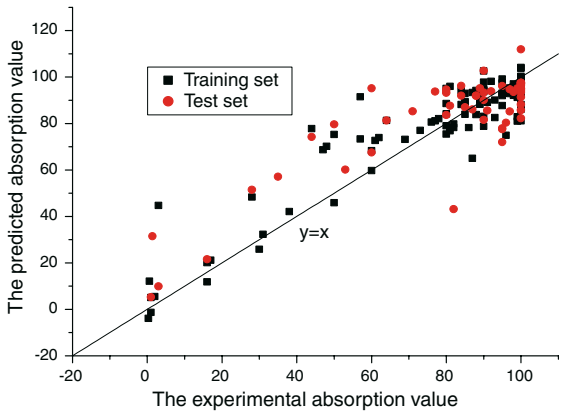

Figure 5. Predicted absorption values vs. experimental values for the training set and test set by support vector machine (C= 100, $\gamma = 0.021, \varepsilon = 0.15$).

Analysis of the results obtained indicates that the model we proposed can correctly represent structure–absorption relationships of these com-

pounds and molecular descriptors calculated solely from structures could describe the structural features of the diffusion rate-limited drugs responsible for their absorption. By comparison of results from the HM and SVM, the performance of SVM model are better than that of the HM method especially for the test set, which indicates that non-linear model can describe the relationship between the structural descriptors and the absorption of drugs more accurately.

Conclusions

The SVM was used to develop the non-linear quantitative structure–absorption relationship of 169 drugs only based on calculated descriptors. The HM was used to select descriptors responsible for absorption of drugs and develop the linear model. Both the linear and non-linear models gave the satisfactory results, at the same time, the non-linear SVM models produced better results with good predictive ability for the test set than the linear model. We can conclude that (1) the proposed models could identify and provide some insight into what structural features are related to the absorption of the diffusion rate-limited drugs. (2) Non-linear relationship can describe accurately the relationship between the structural parameters and the absorption of the 169 drugs. (3) SVM proved to be a useful tool in the prediction of the absorption behavior of the drugs. It has some advantages over the other techniques of converging to the global optimum, and not to a local optimum. Besides, because only support vectors (only a fraction of all data) are used in the generalization process, the SVM adapts particularly to the problems with a great deal of data in cheminformatics. Therefore, the SVM is a very promising machine learning technique from many aspects and will gain more extensive applications.

Acknowledgements

The authors thank the Association Franco–Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Programme PRA SI 02-03). The authors also thank the R Development Core Team for affording the free R1.7.1 software.

References

1. Yang, C.Y., Dantzig, A.H. and Pidgeon, C., Pharm. Res., 16 (1999) 1331.
2. Zhao, Y.H., Le, J.A.M.H., Hersey, A., Eddershaw, P.J., Luscombe, C.N., Boutina, D., Beck, G., Sherborne, B., Cooper, I.J. and Platts, A.J., Pharm. Sci., 90 (2001) 749.
3. Yao, X.J., Liu, M.C., Zhang, X.Y., Hu, Z.D. and Fan, B.T., Anal. Chim. Acta, 462 (2002) 101.
4. Raevsky, O.A., Fetisov, V.I., Trepalina, E.P., McFarland, J.W. and Schaper, K.J., Quant. Struct.-Act. Relat., 19 (2000) 366.
5. Clark, D.E., J. Pharm. Sci., 88 (1999) 807.
6. Wessel, M.D., Jurs, P.C., Tolan, J.W. and Muskal, S.M., J. Chem. Inf. Comput. Sci., 38 (1998) 726.
7. Zhao, Y.H., Abraham, M.H., Le, J., Hersey, A., Luscombe, C.N., Beck, G., Sherborne, B. and Cooper, I., Pharm. Res., 19 (2002)1446.
8. Tavelin, S., Taipalensuu, J., Söderberg, L., Morrison, R., Chong, S. and Artursson, P., Pharm. Res., 20 (2003) 397.
9. Wegner, J.K., Fröhlich, H. and Zell, A., J. Chem. Inf. Comput. Sci., 44 (2004) 931
10. Egan, W.J., Merz, K.M. and Baldwin, J.J., J. Med. Chem., 43 (2000) 3867.
11. Stenberg, P., Norinder, U., Kristina, L. and Artursson, P., J. Med. Chem., 44 (2001) 1927.
12. Bergström, C.A.S., Strafford, M., Lazorova, L., Avdeef, A., Luthman, Kristina and Artursson, Per., J. Med. Chem., 46 (2003) 558.
13. Katritzky, A.R., Lobanov, V.S. and Karelson, M., Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0, 1994.
14. Katritzky, A.R., Lobanov, V.S. and Karelson, M., Chem. Soc. Rev., 24 (1995) 279.
15. Oblak, M., Randic, M. and Solmajer, T., J. Chem. Inf. Comput. Sci., 40 (2000) 994.
16. Katritzky, A.R. and Tatham, D.B., J. Chem. Inf. Comput. Sci., 41 (2001) 1162.
17. Burbidge, R., Trotter, M., Buxton, B. and Holden, S., Comput. Chem., 26 (2001) 14.
18. Manallack, D.T. and Livingstone, D.J., Eur. J. Med. Chem., 34 (1999) 95.
19. Goldberg, D., Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA, 1989.
20. Bao, L. and Sun, Z.R., FEBS Lett., 521 (2002) 109.
21. Belousov, A.I., Verzakov, S.A. and Von Frese J., Chemometr. Intell. Lab. Syst., 64 (2002) 15.
22. Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C., Comb. Chem., 26 (2002) 293.
23. Morris, C.W., Autret, A. and Boddy, L., Ecol. Model., 146 (2001) 57.
24. Minghu S.C.M., Breneman, J.B, Sukumar N., K P.B., Steven C. and Nihal T., J. Chem. Inf. Comput. Sci., 42 (2002) 1347.
25. Liu, H.X., Zhang, R.S., Luan, F., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Chem. Inf. Comput. Sci., 43 (2003) 900.
26. 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.
27. 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.

28. 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.
29. Katritzky, A.R., Petrukhin, R., Jain, R. and Karelson, M., J. Chem. Inf. Comput. Sci., 41 (2001) 1521.
30. Cortes, C. and Vapnik, V., Machine Learning, 20 (1995) 273.
31. Zhang, L., Zhou, W.D. and Jiao, L.C., J. Infrared Millimeter Waves, 21 (2002) 119.
32. Ding, C.H.Q. and Dubchak, I., Bioinformatics, 17 (2001) 349.
33. Karchin, R., Karplus, K. and Haussler, D., Bioinformatics, 18 (2002) 147.
34. Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C., J. Comput. Chem., 23 (2002) 267.
35. Wang, W.J., Xu, Z.B., Lu, W.Z. and Zhang, X.Y., Neurocomputing, 55 (2003) 643.
36. Vapnik, V. Statistical Learning Theory, Wiley, New York, 1998.
37. Schölkopf, B., Burges, C. and Smola, A., Advances in Kernel Methods – Support Vector Learning, MIT Press, Cambridge, MA, 1999.
38. Cristianini, N. and Shawe-Taylor, J. An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.
39. Sietsema, W.K., Int. J. Clin. Pharmacol. Ther. Toxicol., 27 (1989) 179.
40. Bosque, R. and Sales, J., J. Chem. Inf. Comput. Sci., 43 (2003) 637.
41. Katritzky, A.R., Oliferenko, A.A., Oliferenko, P.V., Petrukhin, R. and Tatham, D.B., J. Chem. Inf. Comput. Sci., 43 (2003) 1794.
42. Balon, B., Riebesehl, B.U. and Müller, B.W., Pharm. Res., 16 (1999) 890.