Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 43 (2008) 2442–2452
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

Explorations into modeling human oral bioavailability

Zhi Wang [a], Aixia Yan [a,*], Qipeng Yuan [a], Johann Gasteiger [b,c]

[a] State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, PR China
[b] Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany
[c] Molecular Networks GmbH, Henkestrasse 91, D-91052 Erlangen, Germany

Received 6 March 2008; received in revised form 5 May 2008; accepted 15 May 2008
Available online 28 May 2008

Abstract

Explorations into modeling human oral bioavailability started with a whole dataset of 772 drug compounds. First, training set and test set were chosen based on Kohonen's self-organizing Neural Network (KohNN). Then, a quantitative model of the whole dataset was built using multiple linear regression (MLR) analysis. This model had limited predictability emphasizing that a variety of pharmacokinetic factors influence human oral bioavailability. In order to explore whether better models can be built when the compounds share some ADME properties, four subsets were chosen from the whole dataset to build quantitative models and better models were obtained by MLR analysis. These studies show that, indeed, good models for predicting human oral bioavailability can be obtained from datasets sharing certain pharmacokinetic properties.

Keywords: Quantitative structure–activity relationships (QSAR); Bioavailability; Multiple linear regression (MLR); Leave-one-out cross-validation; Kohonen's self-organizing Neural Network (KohNN)

1. Introduction

In the design of new drugs, one of the most important considerations is to increase the oral bioavailability of a drug candidate. It is reported that 95% of lead compounds fail in the developmental stages, and 50% of these failures were shown to be due to unfavorable absorption, distribution, metabolism, and excretion (ADME) properties [1,2].

Among pharmacokinetic properties, an inadequate bioavailability is one of the main reasons that many promising drug candidates fail before further development. Bioavailability represents the percentage of an oral dose which is able to produce a pharmacological activity, in other words, the fraction of the oral dose that reaches the arterial blood in an active form. Oral bioavailability is related to several factors, such as gastrointestinal transition and absorption, intestinal membrane permeation, and intestinal/hepatic first-pass metabolism. Moreover, in the access of absorption, many authors have suggested that gut wall CYP3A4 and P-glycoprotein act in a concerted manner to control the absorption of their substrates [3–8]. The obvious method to maximize oral absorption would be to design the characteristics of a molecule which make it a substrate of P-glycoprotein and CYP3A4 [9].

Predicting oral bioavailability in silico may be guided by Lipinski's 'Rule of Five', which could be used to predict the absorption and permeability of drug molecules qualitatively [10]. In 1994, Hirono and colleagues [11] reported a study on the quantitative structure–bioavailability relationship for 188 noncongeneric organic drugs; the drugs were separated into three groups: nonaromatics, aromatics, and heteroaromatics, then based on each group a quantitative model was developed. In 2000, Yoshida and colleagues published a classification model for human oral bioavailability [12]. This model can get a correct rate of classification of 60% for the test group using three distribution related descriptors and 15 structure

* Corresponding author. Tel.: +86 10 64421335; fax: +86 10 64416428.
E-mail addresses: aixia_yan@yahoo.com, yanax@mail.buct.edu.cn (A. Yan).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.05.017

descriptors which were considered to be related to some metabolic processes. The descriptors were chosen by analyzing bioavailability in relation to physicochemical and structural factors by the ORMUCS (ordered multicategorical classification method using the simplex technique) [12].

In subsequent work, other researchers have introduced rules-of-thumb which can increase the chances of drug compounds being well absorbed. In 2002, Veber and colleagues reported studies on rat bioavailability data for 1100 drug candidates [13]. It was found that the drug molecules having fewer than 10 rotatable bonds and less than 140 $Å^2$ PSA (polar surface area) (or an H-bond count less than 12) usually showed more than 20% rat oral bioavailability. In 2004, Lu and colleagues investigated the relationship between the number of rotatable bonds and PSA for rat oral bioavailability using 434 molecules [14]. Compared to Veber's work [13], Lu reported that the prediction results were dependent on the calculation methods [14].

In 2007, Hou and colleagues collected a dataset of 773 compounds with experimental human oral bioavailability values [15]. Then, Veber's rules [13] were used for the entire dataset to see if these rules could be applied for the prediction of human oral bioavailability. Afterwards, the correlations between several important molecular descriptors and human oral bioavailability were examined. They conjectured that there are no simple rules based on molecular descriptors that can be used to predict human oral bioavailability truly well compared to the rules based on analyzing rat oral bioavailability data [15].

It is clear that powerful descriptors related to carrier-mediated transport and first-pass metabolism are needed for building a useful prediction model for human oral bioavailability. However, it seems that until now, for a diverse dataset of drug compounds such as the above 773 drug compounds, no set of calculated descriptors was able to represent the complicated relationship between human oral bioavailability and structure.

Hence, in this study, we intended to explore the relationships between human oral bioavailability of drugs and their structures, by building quantitative models from two sides: (1) by including experimental HIA (human intestinal absorption) value as an input descriptor as this property is considered to be related to absorption and metabolism processes; (2) by building individual quantitative models for three different groups of drug compounds, which exhibit similar structures or pharmacological activities.

In this work five different sets of compounds were investigated. (1) Initially, the whole dataset of Hou et al. [15], including 772 drug compounds was studied in Models 1. It was investigated whether a quantitative prediction model could be built based on the 141 available descriptors. Then, four other prediction models were constructed for four different groups. (2) In Models 2, 161 drug compounds, for which experimental HIA values were available, were studied using the HIA value as a descriptor to increase the prediction power of the model. (3) Models 3 and Models 4 were built for specific types of compounds, i.e., for 51 sulfonamide and for 29 β-lactam drug compounds, respectively. (4) Models 5 were constructed based on 58 central nervous system (CNS) drug compounds, which were considered as having similar pharmacological activity. CNS drugs which can restrain or excite the action of central nervous system are considered to have similar absorption and metabolism profiles.

To compare the performance of descriptors from ADRIANA.Code [16] and from those of Cerius^2 [17], individual models were built with the descriptors from ADRIANA.Code, from Cerius^2 and with a combination of them.

2. Methods

2.1. Dataset of human oral bioavailability

In this work, the human oral bioavailability dataset of 772 compounds was taken from http://modem.ucsd.edu/adme, which was collected from 185 literatures, by Hou and colleagues [15]. Experimental human oral bioavailability data is defined as the fraction of the oral dose that reaches the arterial blood in an active form. The original dataset included 773 compounds, however, one compound in this dataset lacked an experimental human oral bioavailability value and was therefore removed. All the structures of the dataset (especially chirality) were checked through Cambridge Chemfinder Database [18] and Chemical Information Specialized Information Services of National Library of Medicine of US [19].

Experimental HIA values for 161 compounds used in the subset for Models 2 were also taken from a dataset collected by Hou et al. [20].

2.2. Molecular descriptors

Two program packages ADRIANA.Code [16] and Cerius^2 [17] were used to calculate the 141 molecular descriptors of this study.

2.2.1. Descriptors calculated by ADRIANA.Code

Seventy-four descriptors including molecular weight (MW), topological polar surface area (TPSA), mean molecular polarizability (MMP) and 2D property-weighted autocorrelation were calculated using ADRIANA.Code [16].

TPSA was calculated using the parameters originally reported by Ertl and colleagues [21]. Mean molecular polarizability (MMP) can be estimated from additive contributions of atoms as shown by Miller [22]. The 2D property-weighted autocorrelation uses the molecule's 2D structure and atom pair properties as a basis for obtaining vectorial molecular descriptors [23,24]. The atom pair properties are summed up for certain topological distances which count the number of bonds on the shortest path between two atoms. Thereby a single value for each topological distance is derived that is one entry in the resulting 2D autocorrelation vector. The 2D molecular autocorrelation vectors [25] were calculated based on the following seven atomic properties: σ charge (SigChg) [26,27], π charge (PiChg), total charges (TotChg), σ electronegativity

(SigEN), π electronegativity (PiEN), lone-pair electronegativity (LpEN) and atomic polarizability (Apolariz) [28].

2.2.2. Descriptors calculated by Cerius^2

Sixty-seven descriptors were calculated using Cerius^2 molecular simulation package, 17 of which include some common descriptors such as octanol–water partitioning coefficient (log P), number of violations of the rule-of-5 ($N_{\text{rule-of-5}}$), rotatable bond count (N_{rot}), H-bond donor count (N_{HBD}), H-bond acceptor count (N_{HAC}), molecular volume (Vm), principal moment of inertia, 10 shadow indices, Jurs descriptors, density (Density), Multigraph information content indices.

The value of log P is related to the hydrophobic character of the molecule. The value of $N_{\text{rule-of-5}}$ is defined as the number of violations of the four rule-of-5 rules proposed by Lipinski et al. [10]. N_{rot} was computed by counting the number of bonds which are allowed to rotate in molecular mechanics (all terminal H-atoms are ignored). The 10 shadow indices which can help to characterize the shape of a molecule were calculated by projecting the molecular surface on three mutually perpendicular planes, *XY*, *YZ*, and *XZ* after rotating the molecules to align the principal moments of inertia with the *X*, *Y*, and *Z* axes [29]. Thirty different descriptors are included in the Jurs descriptor set, which combines shape and electronic information to characterize molecules. The descriptors are computed by mapping atomic partial charges on solvent-accessible surface areas of individual atoms [30]. Information content (IC) and structure information content (SIC = IC/(number of vertices)) are included in Multigraph information content indices.

2.3. Prediction models for bioavailability

Descriptors used in each model were selected according to the correlation between descriptors and bioavailability. The set of descriptors was chosen using stepwise linear regression variable selection method. Stepwise variable entry and removal examines the variables in the block at each step for entry or removal (criteria: probability of F to enter ≤ 0.50, probability of F to remove ≥ 0.100).

2.3.1. Models 1

The whole dataset including 772 compounds was studied first. The SONNIA software [31,32] was used for separating the dataset into a training set and a test set based on Kohonen's self-organizing Neural Network (KohNN). The Kohonen's self-organizing Neural Network has the special property of effectively creating a spatially organized internal representation of various features of input signals and their abstractions [33]. The perception of similarity of objects is an essential feature. In a self-organizing neural network the neurons are arranged in a two-dimensional array to generate a two-dimensional feature map such that similarity in the data is preserved. In other words, if two input data vectors are similar, they will be mapped into the same neuron or closely together in the two-dimensional map. A Kohonen's self-organizing Neural Network was applied to split the dataset into a training set and a test set. This method of splitting a dataset into training set and test set assures that both sets cover the information space as good as possible.

The 772 compounds were split into a training set of 516 compounds and a test set of 256 compounds after KohNN classification. Then MLR (multiple linear regression) analysis was performed to build a set of three quantitative models with the descriptors from ADRIANA.Code (Model 1A), from Cerius^2 (Model 1B) and with a combination of them (Model 1C), respectively.

2.3.2. Models 2

From the 772 compounds, 161 compounds for which experimental HIA values were available were investigated by a Kohonen's self-organizing Neural Network using the SONNIA. This allowed the 161 compounds to be divided into a training set of 125 compounds and a test set of 36 compounds after the KohNN classification. Then MLR (multiple linear regression) analysis was used to build a set of three quantitative models with the descriptors from ADRIANA.Code (Model 2A), from Cerius^2 (Model 2B) and with a combination of them (Model 2C), respectively.

2.3.3. Models 3

From the entire dataset of 772 compounds, all 51 sulfonamide drug compounds were chosen for building a set of three quantitative models with the descriptors from ADRIANA.Code (Model 3A), from Cerius^2 (Model 3B) and with a combination of them (Model 3C), respectively. The regression models were obtained by using MLR analysis with leave-one-out cross-validation method.

2.3.4. Models 4

From the entire dataset of 772 compounds, all 29 β-lactam drug compounds were chosen for building a set of three quantitative models with the descriptors from ADRIANA.Code (Model 4A), from Cerius^2 (Model 4B) and with a combination of them (Model 4C), respectively. The regression models were obtained by using MLR analysis with leave-one-out cross-validation method.

2.3.5. Models 5

From the entire dataset of 772 compounds, all 58 central nervous system (CNS) drug compounds were chosen for building a set of three quantitative models with the descriptors from ADRIANA.Code (Model 5A), from Cerius^2 (Model 5B) and with a combination of them (Model 5C), respectively. The regression models were constructed by using MLR analysis with leave-one-out cross-validation method.

3. Results and discussion

3.1. Models 1

In this set of models, seven descriptors of ADRIANA.Code were selected for building Model 1A; seven descriptors of Cerius^2 were selected for building Model 1B; and 11 descriptors

were selected from the combination for building Model 1C. For the whole dataset of 772 compounds, a rectangular KohNN with 28 × 28 neurons is utilized with 11 combined descriptors (calculated from ADRIANA.Code and Cerius^2) that are used as input vectors. The intercorrelations between the 11 descriptors and bioavailability are given in Table 1.

A rectangular KohNN with 28 × 28 neurons was utilized with the 11 descriptors that are used as input vectors. The initial learning spans are 14 and 14, with an initial learning rate of 0.7 and a rate factor of 0.95. The initial weights are randomly initialized, and training was performed for a period of 1600 epochs in an unsupervised manner. A map was formed according to the ranges of bioavailability of the most frequently occupied neuron. From Fig. 1, one can see that compounds with a different range of bioavailability are projected into different areas. The correct rate of classification is 62.8%, which is calculated as $1 - N_{\text{conflicts}}/N_{\text{total\_compounds}}$ ($N_{\text{conflicts}}$: the number of neurons occupied by compounds which belong to different classes; $N_{\text{total\_compounds}}$: the number of total compounds used in KohNN).

In Kohonen's map, 516 of a total of 784 neurons are occupied. One object of each occupied neuron was taken for the training set; the other objects represented the test set. Thus, the 772 compounds were divided into a training set of 516 compounds and a test set of 256 compounds after the KohNN classification.

MLR (multiple linear regression) analysis was performed with seven descriptors from ADRIANA.Code, seven descriptors from Cerius^2 and 11 descriptors selected from the combination of them for building the corresponding models Model 1A, Model 1B and Model 1C, respectively. The 516 compounds in the training set were used to build a model, and the 256 compounds were used for the prediction of bioavailability.

The following equation was obtained:

$$\text{Bioavailability} = \sum(c_i D_i) + D_c \quad (1)$$

In the equation, D_c is a constant, D_i is a descriptor and c_i is its corresponding regression coefficient in an MLR model (Model 1). The corresponding regression coefficients are shown in Table 2.

For the training set of Model 1A, $r = 0.38$, sd = 31.61, $n = 516$ and for the test set of Model 1A, $r = 0.37$, sd = 30.05 and $n = 256$. For the training set of Model 1B, $r = 0.40$, sd = 31.40, $n = 516$ and for the test set of Model 1B, $r = 0.36$, sd = 30.11 and $n = 256$. For the training set of Model 1C, $r = 0.42$, sd = 31.15, $n = 516$, and $F = 9.85$ and for the test set of Model 1C, $r = 0.43$, sd = 29.20 and $n = 256$ (r is the correlation coefficient and sd is the standard deviation).

Comparing the above three models, one can see similar results obtained by seven descriptors of ADRIANA.Code (Model 1A), seven descriptors of Cerius^2 (Model 1B) and 11 descriptors (Model 1C) selected from the combined descriptor set. It is obvious that these three models obtained are all of rather low predictability. Thus, the full dataset of 772 compounds does not lead to a good model as also observed by Hou and colleagues [15] in their publication.

It is not too surprising that no good prediction model of general validity could be obtained. The factors influencing human oral bioavailability are just too varied including different pharmacokinetic factors such as absorption, distribution and metabolism processes. For the different types of compounds, these factors have different importance and contribute with different weight.

One factor strongly contributing to human oral bioavailability is human intestinal absorption (HIA). Therefore, in Models 2, it was investigated whether an improved prediction model for human oral bioavailability could be obtained by using HIA values as a descriptor.

Next, two datasets of compounds, sulfonamides and β-lactams, were investigated for it can be assumed that compounds of a specific class have similar pharmacokinetic properties (Models 3 and Models 4).

Table 1
The intercorrelations between the 11 descriptors and bioavailability data for 772 drug compounds of the whole dataset

	$N_{\text{rule-of-5}}$	IC	LUMO	Jurs-RPCG	N_{rot}	Log P	TPSA	Acorr_SigChg_3	Acorr_SigChg_5	Acorr_SigChg_7	Acorr_TotChg_4
$N_{\text{rule-of-5}}$[a,b]	1										
IC[b]	0.210	1									
LUMO[b]	−0.039	−0.132	1								
Jurs-RPCG[b]	−0.344	−0.418	−0.114	1							
N_{rot}[a,b]	0.579	0.287	0.017	−0.463	1						
Log P[b]	−0.168	0.005	0.088	−0.147	−0.188	1					
TPSA[a]	0.647	0.279	−0.095	−0.225	0.606	−0.613	1				
Acorr_SigChg_3[a]	−0.290	−0.055	−0.141	0.311	−0.239	−0.127	−0.079	1			
Acorr_SigChg_5[a]	−0.147	−0.032	−0.111	0.059	−0.247	0.141	−0.281	0.249	1		
Acorr_SigChg_7[a]	0.092	0.077	−0.035	−0.091	0.165	−0.070	0.258	0.085	−0.058	1	
Acorr_TotChg_4[a]	0.287	−0.004	0.144	−0.198	0.314	−0.020	0.190	−0.677	−0.690	−0.003	1
Bioavailability	−0.292	−0.163	0.070	0.238	−0.310	0.141	−0.236	0.139	0.069	−0.112	−0.208

$N_{\text{rule-of-5}}$: number of violations of the rule-of-5; IC: information content; LUMO: lowest unoccupied molecular orbital energy; Jurs-RPCG: relative positive charge: charge of most positive atom divided by the total positive charge; N_{rot}: number of bonds which are allowed to rotate in molecular mechanics; log P: octanol−water partitioning coefficient; TPSA: topological polar surface area; Acorr_SigChg_3,5,7: the third, fifth, seventh components of 2D autocorrelation coefficients for σ charge; Acorr_TotChg_4: 2D autocorrelation coefficients for the total charge on the fourth atom.

[a] Descriptors calculated by ADRIANA.Code.

[b] Descriptors calculated by Cerius^2.

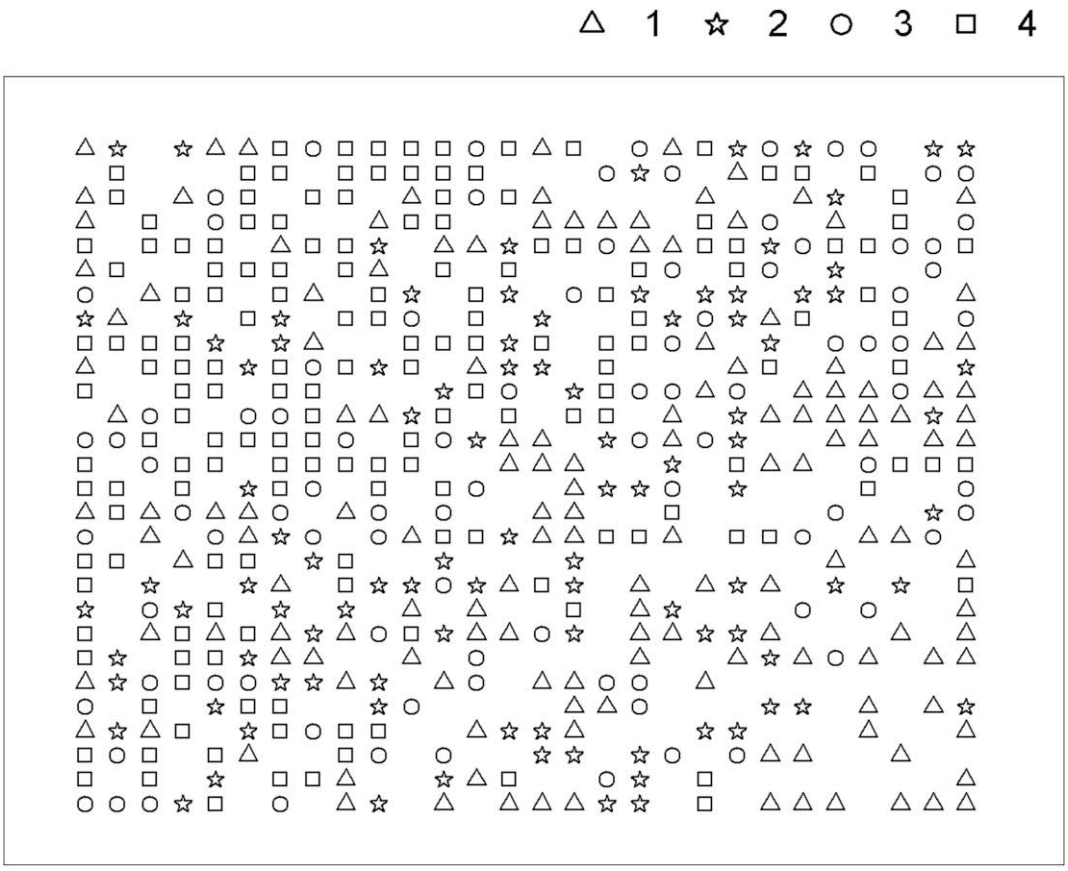

Fig. 1. A rectangular KohNN map for 772 compounds obtained by using the 11 selected input descriptors; '1' means compounds with bioavailability in the range of 0–25%, '2' means compounds with bioavailability in the range of 25–50%, '3' refers to compounds with bioavailability in the range of 50–75%, and '4' refers to compounds with bioavailability in the range of 75–100%.

Table 2
Descriptors selected in each model and their corresponding regression coefficients in the multiple linear regression for Models 1 based on the whole dataset of 772 drug compounds

Model 1A		Model 1B		Model 1C	
Descriptors	Coefficients, D_i	Descriptors	Coefficients, D_i	Descriptors	Coefficients, D_i
TPSA	0.20	IC	−4.30	$N_{\text{rule-of-5}}$[b]	−10.41
Acorr_SigChg_1	−39.15	LUMO	1.93	IC[b]	1.50
Acorr_SigChg_3	−17.18	Jurs-FPSA-3	−197.33	LUMO[b]	−6.60
Acorr_SigChg_5	−10.95	Jurs-RPCG	67.02	Jurs-RPCG[b]	68.47
Acorr_SigChg_7	−10.01	$N_{\text{rule-of-5}}$	−6.76	N_{rot}[b]	−1.31
Acorr_TotChg_4	−15.79	N_{rot}	−0.91	Log P[b]	1.67
Acorr_PiEN_4	0.01	Log P	0.62	TPSA[a]	0.12
				Acorr_SigChg_3[a]	−10.22
				Acorr_SigChg_5[a]	−9.94
				Acorr_SigChg_7[a]	−9.30
				Acorr_TotChg_4[a]	−17.88
D_c	63.042	D_c	87.688	D_c	64.51

$N_{\text{rule-of-5}}$: number of violations of the rule-of-5; IC: information content; LUMO: lowest unoccupied molecular orbital energy; Jurs-RPCG: relative positive charge: charge of most positive atom divided by the total positive charge; Jurs-FPSA-3: the third descriptor of fractional charged partial surface areas; N_{rot}: number of bonds which are allowed to rotate in molecular mechanics; log P: octanol–water partitioning coefficient; TPSA: topological polar surface area; Acorr_SigChg_1,3,5,7: the first, third, fifth, seventh components of 2D autocorrelation coefficients for σ charge; Acorr_PiChg_8: the eighth component of 2D autocorrelation coefficients for π charge; Acorr_TotChg_4: 2D autocorrelation coefficients for the total charge on the fourth atom; Acorr_PiEN_4: the fourth component of 2D autocorrelation coefficients for π electronegativity.

Model 1A was based on seven selected ADRIANA.Code descriptors, Model 1B was based on seven selected Cerius[2] descriptors and Model 1C was based on 11 combined descriptors.

[a] Descriptors calculated by ADRIANA.Code.

[b] Descriptors calculated by Cerius[2].

Finally, a dataset was investigated that comprised compounds having the same location of exerting its biological activity, i.e., all compounds are being active in the central nervous system (CNS). For, it can be assumed that such compounds have to undergo similar absorption and distribution processes (Models 5).

3.2. Models 2

In this set of models, four descriptors of ADRIANA.Code were selected for building Model 2A; four descriptors of Cerius[2] were selected for building Model 2B; and six descriptors were selected from the combination for building Model 2C. For each model, the experimental HIA value was added as an additional descriptor. For the 161 compounds which have experimental HIA data, a rectangular KohNN with 14×13 neurons is utilized with seven combined descriptors that are used as input vectors. The intercorrelations between seven descriptors and bioavailability are given in Table 3.

A rectangular KohNN with 14×13 neurons was utilized with the seven descriptors that are used as input vectors. The initial learning spans are 7 and 6.5, with an initial learning rate of 0.7 and a rate factor of 0.95. The initial weights are randomly initialized, and the training was performed for a period of 1600 epochs in an unsupervised manner. A map was formed according to the ranges of bioavailability of the most frequently occupied neuron. From Fig. 2, one can see that compounds with a different range of bioavailability are projected into different areas. The correct rate of classification is 78.9%.

In the Kohonen's map, 125 of a total of 182 neurons are occupied. One object from an occupied neuron was taken for the training set, the other objects represented the test set. Thus, the 161 compounds were divided into a training set of 125 compounds and a test set of 36 compounds after the KohNN classification.

Three MLR analyses were performed using four descriptors from ADRIANA.Code and HIA, four descriptors from Cerius[2] and HIA, and seven combined descriptors for building corresponding models Model 2A, Model 2B and Model 2C, respectively. The 125 compounds in the training set were used to build a model, and the 36 compounds were used for the prediction of bioavailability.

The following equation has been obtained:

$$\text{Bioavailability} = \sum (c_i D_i) + D_c \quad (2)$$

In the equation, D_c is a constant, D_i is a descriptor and c_i is its corresponding regression coefficient in an MLR model (Models 2). The corresponding regression coefficients are shown in Table 4.

For the training set of Model 2A, $r = 0.58$, $\text{sd} = 29.02$, $n = 125$, and for the test set of Model 2A, $r = 0.52$, $\text{sd} = 30.41$, and $n = 36$. For the training set of Model 2B, $r = 0.56$, $\text{sd} = 29.51$, $n = 125$, and for the test set of Model 2B, $r = 0.39$, $\text{sd} = 32.80$, and $n = 36$. For the training set of Model 2C, $r = 0.62$, $\text{sd} = 28.25$, $n = 125$, and $F = 10.49$ and for the test set of Model 2C, $r = 0.57$, $\text{sd} = 29.35$, and $n = 36$ (r is the correlation coefficient and sd is the standard deviation). The results of Model 2C are shown in Fig. 3.

Comparing the above three models, one can see similar results obtained by Model 2A and Model 2B; Model 2C building with seven combined descriptors seems slightly better than Model 2A and Model 2B.

Before HIA was added as a descriptor, to compare with Model 2C, another regression model named Model 2D was built based on the other six descriptors. The comparison of the two models is shown in Table 5.

Model 2C is slightly better than Models 1, but it is still not a good prediction model because the factors influencing human oral bioavailability are complicated. The purpose of building Models 2 (four models) is to see whether HIA (human intestinal absorption) could be used to increase the quality of prediction model. By comparing Model 2C and Model 2D, it could be observed that HIA is really an important factor that influences the human oral bioavailability.

For the training set of Model 2C, $r = 0.62$, for Model 2D, $r = 0.45$; for the test set of Model 2C, $r = 0.57$, for Model 2D, $r = 0.49$. The RMS deviations for the two models are 27.68 and 30.97, respectively. From these results, it is obvious that HIA is a useful descriptor, as it leads to a remarkable

Table 3
The intercorrelations between the seven descriptors and bioavailability for 161 drug compounds with HIA experimental value

	HIA	IC	Log *P*	Acorr_PiChg_2	Acorr_LpEN_1	TPSA	MMP
HIA[c]	1						
IC[b]	−0.007	1					
Log *P*[b]	0.621	0.023	1				
Acorr_PiChg_2[a]	−0.123	−0.038	0.180	1			
Acorr_LpEN_1[a]	−0.313	0.271	−0.472	−0.128	1		
TPSA[a]	−0.711	0.218	−0.782	−0.113	0.528	1	
MMP[a]	−0.096	0.574	0.103	0.078	0.120	0.277	1
Bioavailability	0.498	−0.131	0.225	−0.259	0.011	−0.245	−0.175

IC: information content; HIA: human intestinal absorption; TPSA: topological polar surface area; MMP: mean molecular polarizability; log *P*: octanol−water partitioning coefficient; Acorr_PiChg_2: the second component of 2D autocorrelation coefficients for π charge; Acorr_LpEN_1: the first component of 2D autocorrelation coefficients for lone-pair electronegativity.

[a] Descriptors calculated by ADRIANA.Code.

[b] Descriptors calculated by Cerius[2].

[c] Descriptors with experimental value.

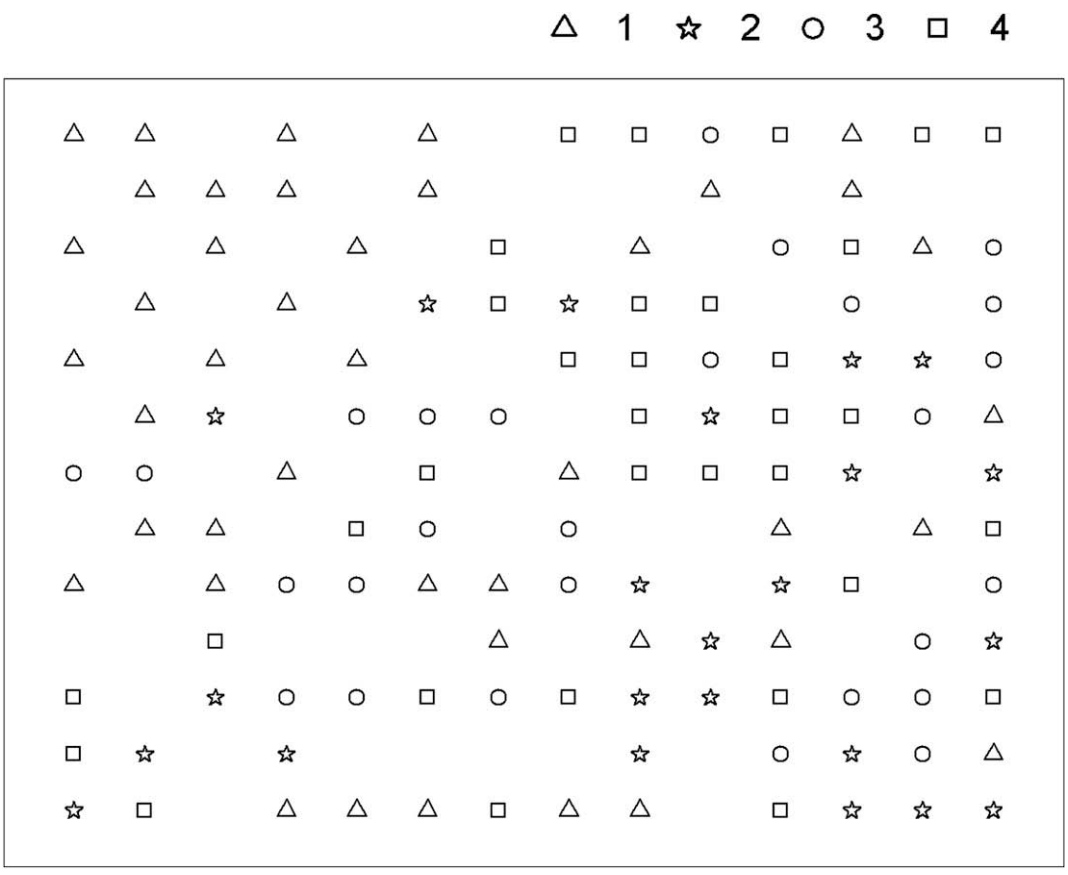

Fig. 2. A rectangular KohNN map for 161 compounds obtained by using the seven combined descriptors as input descriptors; '1' means compounds with bioavailability in the range of 0–25%, '2' means compounds with bioavailability in the range of 25–50%, '3' refers to compounds with bioavailability in the range of 50–75%, and '4' refers to compounds with bioavailability in the range of 75–100%.

increase in the correlation coefficient. It reveals that HIA can be used for predicting human oral bioavailability.

However, in the design of new drugs, HIA is also a difficult pharmacokinetic property which needs careful experiments for measuring it. Thus, experimental values on HIA are available for only a limited set of compounds. In this situation, recourse has to be made for values of HIA predicted by a computational model. This is particularly needed for libraries of virtual

Table 4
Descriptors selected in each model and their corresponding regression coefficients in the multiple linear regression for Models 2 based on 161 drug compounds with HIA experimental value

Model 2A		Model 2B		Model 2C	
Descriptors	Coefficients, D_i	Descriptors	Coefficients, D_i	Descriptors	Coefficients, D_i
HIA[c]	0.80	HIA[c]	0.66	HIA[c]	0.76
Acorr_PiChg_2	−8.25	N_{rot}	−1.49	IC[b]	−14.41
Acorr_LpEN_1	0.03	N_{HBD}	2.00	Log P[b]	1.85
TPSA	0.13	IC	−9.39	Acorr_PiChg_2[a]	−56.07
MMP	−0.57	Log P	−0.89	Acorr_LpEN_1[a]	0.04
				TPSA[a]	0.24
				MMP[a]	−0.46
D_c	−9.13	D_c	31.42	D_c	29.01

IC: information content; HIA: human intestinal absorption; TPSA: topological polar surface area; MMP: mean molecular polarizability; log P: octanol–water partitioning coefficient; N_{rot}: number of bonds which are allowed to rotate in molecular mechanics; N_{HBD}: number of H-bond donor count; Acorr_PiChg_2: the second component of 2D autocorrelation coefficients for π charge; Acorr_LpEN_1: the first component of 2D autocorrelation coefficients for lone-pair electronegativity.

Model 2A was based on four selected ADRIANA.Code descriptors and HIA, Model 2B was based on four selected Cerius^2 descriptors and HIA, and Model 2C was based on seven combined descriptors.

[a] Descriptors calculated by ADRIANA.Code.

[b] Descriptors calculated by Cerius^2.

[c] Descriptors with experimental value.

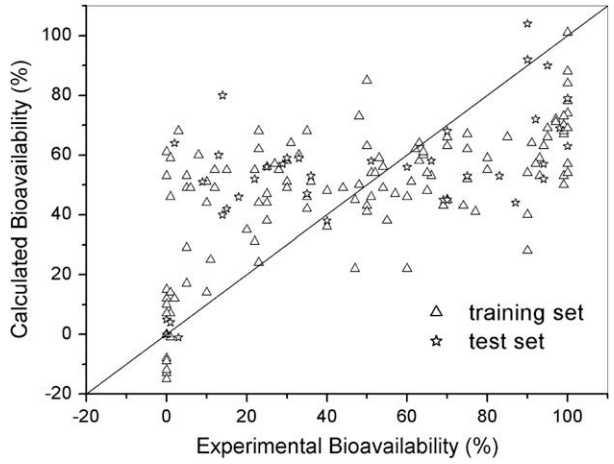

Fig. 3. Calculated vs. experimental values of bioavailability for the 161 drug compounds in Model 2C based on seven combined descriptors (including experimental HIA value as a descriptor) by multiple linear regression analysis.

compounds. Therefore, work by our group on the development of a model for predicting human intestinal absorption (our unpublished results) is in progress.

3.3. Models 3

Fifty-one sulfonamide drug compounds which can be considered having similar metabolism pathways were chosen for building models. Three MLR analyses with leave-one-out cross-validation method were performed using four descriptors from ADRIANA.Code (Model 3A), four descriptors of Cerius^2 (Model 3B), and six combined descriptors (Model 3C) for building corresponding models, respectively. Descriptors selected in each model and the results of MLR analyses with the leave-one-out cross-validation method are shown in Table 6.

For Model 3A, $r = 0.68$, sd $= 20.81$, $n = 51$; for Model 3B, $r = 0.44$, sd $= 30.45$, $n = 51$; and for Model 3C, $r = 0.77$, sd $= 21.66$, $n = 51$. It can be observed that the performance of Model 3A built with four ADRIANA.Code descriptors is better than that of Model 3B built with four Cerius^2 descriptors. The performance of Model 3C built with six combined descriptors is still better than the other two models. The results of Model 3C are shown in Fig. 4.

3.4. Models 4

Twenty-nine β-lactam drug compounds which can be considered having similar absorption and metabolism pathways were chosen for building models. Three MLR analyses with leave-one-out cross-validation method were performed using five descriptors from ADRIANA.Code (Model 4A), five descriptors from Cerius^2 (Model 4B), and five combined descriptors (Model 4C) for building corresponding models, respectively. Descriptors selected in each model and the results of MLR analyses with leave-one-out cross-validation method are shown in Table 7.

For Model 4A, $r = 0.78$, sd $= 23.98$, $n = 29$; for Model 4B, $r = 0.69$, sd $= 27.89$, $n = 29$; and for Model 4C, $r = 0.78$, sd $= 23.98$, $n = 29$. In this set of models, the performance of Model 4A built with five ADRIANA.Code descriptors is

Table 5
Model 2C is based on the selected seven descriptors above (including experimental HIA value as a descriptor), while Model 2D is based on the other six descriptors without HIA

Model	Training set			Test set			RMS
	n	*r*	sd	*n*	*r*	sd	
Model 2C	125	0.62	28.25	36	0.57	29.35	27.68
Model 2D	125	0.45	32.05	36	0.49	30.95	30.97

Table 6
Descriptors selected in each model and the results of MLR analyses with leave-one-out cross-validation method for Models 3 based on 51 sulfonamide drug compounds

Model 3A	Model 3B	Model 3C
Descriptors	Descriptors	Descriptors
Acorr_TotChg_6	Log *P*	Log *P*[b]
Acorr_LpEN_9	SIC	SIC[b]
Acorr_PiEN_8	N_{HAC}	Acorr_TotChg_6[a]
Acorr_LpEN_4	Shadow-*YZ*frac	Acorr_LpEN_9[a]
		Acorr_PiEN_8[a]
		Acorr_LpEN_4[a]
$r = 0.68$ sd $= 20.81$	$r = 0.44$ sd $= 30.45$	$r = 0.77$ sd $= 21.66$

N_{HAC}: number of H-bond acceptor count; shadow-*YZ*frac: fraction of area of molecular shadow in the *YZ* plane over area of enclosing rectangle; SIC: structural information content; Acorr_TotChg_6: the sixth component of 2D autocorrelation coefficients for the total charge; Acorr_LpEN_9: the ninth component of 2D autocorrelation coefficients for lone-pair electronegativity; Acorr_PiEN_8: the eighth component of 2D autocorrelation coefficients for π electronegativity; Acorr_LpEN_4: the fourth component of 2D autocorrelation coefficients for lone-pair electronegativity.
Model 3A was based on four selected ADRIANA.Code descriptors, Model 3B was based on four selected Cerius^2 descriptors, and Model 3C was based on six combined descriptors.
[a] Descriptors calculated by ADRIANA.Code.
[b] Descriptors calculated by Cerius^2.

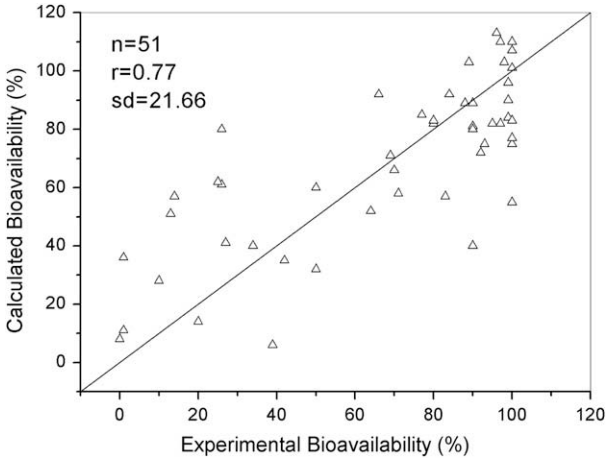

Fig. 4. Calculated vs. experimental values of bioavailability for the 51 sulfonamide drug compounds in Model 3C by multiple linear regression analysis with leave-one-out cross-validation method.

Table 7
Descriptors selected in each model and the results of MLR analyses with leave-one-out cross-validation method for Models 4 based on 29 β-lactam drug compounds

Model 4A	Model 4B	Model 4C
Descriptors	Descriptors	Descriptors
Acorr_SigChg_4	Shadow-*Y* length	Acorr_SigChg_4[a]
Acorr_LpEN_9	N_{cc}	Acorr_LpEN_9[a]
Acorr_TotChg_4	N_{HBD}	Acorr_TotChg_4[a]
Acorr_PiChg_7	Jurs-RPCS	Acorr_PiChg_7[a]
Acorr_SigChg_5	Jurs-DPSA-3	Acorr_SigChg_5[a]
$r = 0.78$ sd = 23.98	$r = 0.69$ sd = 27.89	$r = 0.78$ sd = 23.98

N_{HBD}: number of H-bond donor count; N_{cc}: number of chirality center; Jurs-RPCS: relative positive charge surface area: solvent-accessible surface area of the most positive atom divided by descriptor Jurs-RPCG; Jurs-DPSA-3: difference in atomic charge weighted surface areas; Shadow-*Y* length: length of molecule in the *Y* dimension; Acorr_SigChg_4: the fourth component of 2D autocorrelation coefficients for σ charge; Acorr_LpEN_9: the ninth component of 2D autocorrelation coefficients for lone-pair electronegativity; Acorr_TotChg_4: the fourth component of 2D autocorrelation coefficients for the total charge; Acorr_PiChg_7: the seventh component of 2D autocorrelation coefficients for π charge; Acorr_SigChg_5: the fifth component of 2D autocorrelation coefficients for σ charge.
Model 4A was based on five selected ADRIANA.Code descriptors, Model 4B was based on five selected Cerius^2 descriptors, and Model 4C was based on five combined descriptors.
[a] Descriptors calculated by ADRIANA.Code.

slightly better than that of Model 4B built with five Cerius^2 descriptors; Model 4C built with the five combined descriptors is as same as Model 4A. The results of Model 4C are shown in Fig. 5.

It can be seen that both Models 3 and Models 4 have good results. It is because these two datasets of compounds, sulfonamides (for building Models 3) and β-lactams (for building Models 4) that were investigated include same analog compounds of a specific class with similar pharmacokinetic properties.

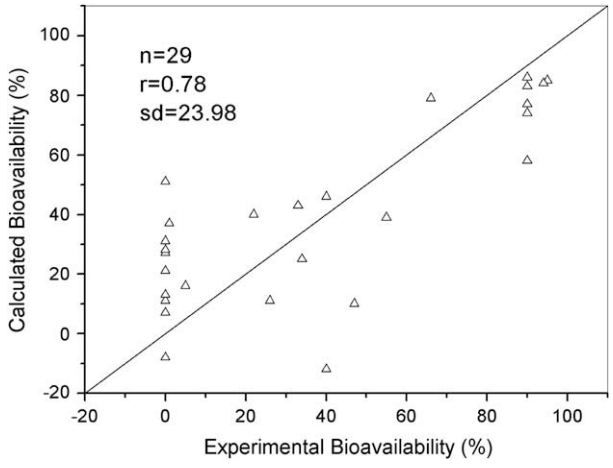

Fig. 5. Calculated vs. experimental values of bioavailability for the 29 β-lactam drug compounds in Model 4C by multiple linear regression analysis with leave-one-out cross-validation method.

3.5. Models 5

Fifty-eight central nervous system drug compounds were chosen for building models. Three MLR analyses with leave-one-out cross-validation method were performed using four descriptors from ADRIANA.Code (Model 5A), four descriptors from Cerius^2 (Model 5B), and seven combined descriptors (Model 5C) for building corresponding models, respectively. Descriptors selected in each model and the results of MLR analyses with leave-one-out cross-validation method are shown in Table 8.

For Model 5A, $r = 0.79$, sd = 19.26, $n = 58$; for Model 5B, $r = 0.70$, sd = 22.50, $n = 58$; and for Model 5C, $r = 0.80$, sd = 18.72, $n = 58$. It can be seen that the performance of Model 5A built with four ADRIANA.Code descriptors is slightly better than that of Model 5B built with four Cerius^2 descriptors, and the performance of Model 5C built with seven combined descriptors is nearly the same as Model 5A and slightly better than Model 5B. The results of Model 5C are shown in Fig. 6.

Drug compounds in Models 5 are CNS drugs which are active in the central nervous system. Good models were obtained for CNS drugs belonging to the same pharmacological activity, which can be thought of having similar absorption and metabolism mechanisms.

4. Conclusions

In this work, first, a diverse dataset of drug compounds with human oral bioavailability values was studied to obtain a regression model. From the results, one can see that the model

Table 8
Descriptors selected in each model and the results of MLR analyses with leave-one-out cross-validation method for Model 5 based on 58 central nervous system drug compounds

Model 5A	Model 5B	Model 5C
Descriptors	Descriptors	Descriptors
Acorr_Ident_3	Shadow-*XY*	Jurs-FPSA-1[b]
Acorr_TotChg_2	Jurs-RPCG	Jurs-DPSA-1[b]
Acorr_Ident_7	Jurs-FPSA-1	Shadow-*XY*[b]
Acorr_PiEN_9	Jurs-DPSA-1	Acorr_Ident_3[a]
		Acorr_TotChg_2[a]
		Acorr_Ident_7[a]
		Acorr_PiEN_9[a]
$r = 0.79$ sd = 19.26	$r = 0.70$ sd = 22.50	$r = 0.80$ sd = 18.72

Jurs-RPCG: relative positive charge: charge of most positive atom divided by the total positive charge; Jurs-FPSA-1: the first descriptor of fractional charged partial surface areas; Jurs-DPSA-1: difference in charged partial surface areas; Shadow-*XY*: area of the molecular shadow in the *XY* plane; Acorr_Ident_3: the third component of 2D autocorrelation coefficients for the atom identity; Acorr_TotChg_2: the second component of 2D autocorrelation coefficients for the total charge; Acorr_Ident_7: the seventh component of 2D autocorrelation coefficients for the atom identity; Acorr_PiEN_9: the ninth component of 2D autocorrelation coefficients for π electronegativity.
Model 5A was based on four selected ADRIANA.Code descriptors, Model 5B was based on four selected Cerius^2 descriptors, and Model 5C was based on seven combined descriptors.
[a] Descriptors calculated by ADRIANA.Code.
[b] Descriptors calculated by Cerius^2.

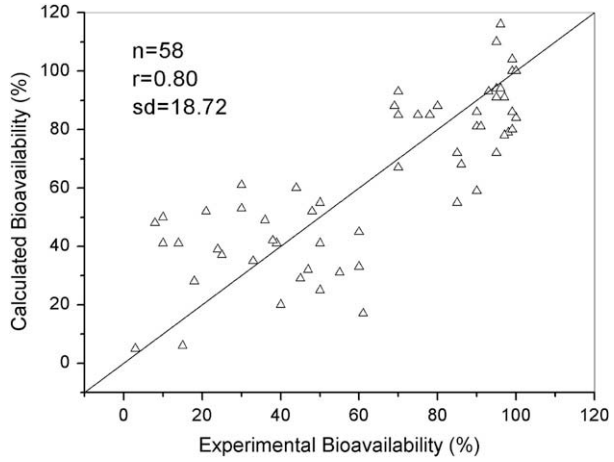

Fig. 6. Calculated vs. experimental values of bioavailability for the 58 central nervous system drug compounds in Model 5C by multiple linear regression analysis with leave-one-out cross-validation method.

obtained is of limited predictability. This is in agreement with work by other groups [15]. The pharmacokinetic factors influencing human oral bioavailability are just too diversified and exert these influences with different weights to an extent that no general model for the prediction of human oral bioavailability for a diverse set of compounds might be achievable. In order to explore whether prediction models for some limited datasets can be obtained, four subsets of the dataset were chosen and four regression models for human oral bioavailability were constructed. First, experimental human intestinal absorption (HIA) which is considered including some distribution and metabolism properties was used as a descriptor. Clearly, HIA is an important descriptor leading to a better model than that without the HIA descriptor. However, as experimental data of HIA are still difficult to obtain, work has to be devoted to obtain a prediction model for HIA with the resulting computed HIA values then to be used for the prediction of human oral bioavailability.

Furthermore, it was found that better MLR models could be obtained for subsets of drugs, which have similar structure, as shown by Models 3 and Models 4 in this work. From Models 5 developed for compounds active in the central nervous system, one can see that it is also worth to build regression models for drug compounds which belong to the same pharmacological activity, which can be thought of having similar absorption and metabolism mechanisms. Subsets used in this study (for building Models 3–Models5) were manually chosen because they have obvious similar structure or pharmacological activity. We just chose the subset that ensures the number of compounds is large enough to guarantee the significance of the statistical analysis. Among the rest of the drugs of this 773 database, subsets of drugs which have similar structure still exist, such as antipyretic and analgesics, but the number of drugs in these subsets is little. In addition, there may be other subsets of drugs with the same pharmacological activity that need to be found, such as circulatory system agents.

Descriptors calculated by different softwares (ADRIANA.Code and Cerius^2) were also compared in this study. From the five sets of models of this study, it can be seen that the performances of models built with ADRIANA.Code descriptors are slightly better than those of models built with Cerius^2 descriptors.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (20605003), National High Tech Project (2006AA02Z337), SRF for ROCS, and the ‘‘Special Funding for the Talent Enrollment’’ of Beijing University of Chemical Technology.

The authors thank Dr. Tingjun Hou and the ADME research team of Department of Chemistry and Biochemistry, UCSD for their database on bioavailability. We thank Molecular Networks GmbH, Erlangen, Germany for making the programs ADRIANA.Code and SONNIA available for our scientific work. We are also grateful to Dr. Thomas Kleinoeder of Molecular Networks for his support and helpful comments.

Appendix. Supplementary material

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.ejmech.2008.05.017.

References

[1] A.P. Beresford, H.E. Selick, M.H. Tarbit, Drug Discov. Today 7 (2002) 109–116.
[2] W.J. Egan, K.M. Merz, J.J. Baldwin, J. Med. Chem. 43 (2000) 3867–3877.
[3] J. Van Asperen, O. Van Tellingen, J.H. Beijnen, Pharm. Res. 37 (1998) 429–435.
[4] A. Lampen, Y. Zhang, I. Hackbarth, L.Z. Benet, K.F. Sewing, U. Christians, J. Pharmacol. Exp. Ther. 285 (1998) 1104–1112.
[5] V.J. Wacher, J.A. Silverman, Y. Zhang, L.Z. Benet, J. Pharm. Sci. 87 (1998) 1322–1330.
[6] M.F. Hebert, Adv. Drug Deliv. Rev. 27 (1997) 201–214.
[7] S.D. Hall, K.E. Thummel, P.B. Watkins, K.S. Lown, L.Z. Benet, M.F. Paine, R.R. Mayo, D.K. Turgeon, D.G. Bailey, R.J. Fontana, S.A. Wrighton, Drug Metab. Dispos. 27 (1999) 161–166.
[8] K.S. Lown, R.R. Mayo, A.B. Leichtman, H. Hsiao, D.K. Turgeon, P. Schmiedlin-Ren, M.B. Brown, W. Guo, S.J. Rossi, L.Z. Benet, P.B. Watkins, Clin. Pharmacol. Ther. 62 (1997) 248–260.
[9] H. van de Waterbeemd, D.A. Smith, K. Beaumont, D.K. Walker, J. Med. Chem. 44 (2001) 1313–1333.
[10] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Adv. Drug Deliv. 23 (1997) 3–25.
[11] S. Hirono, I. Nakagome, H. Hirano, Y. Matsushita, F. Yoshi, I. Moriguchi, Biol. Pharm. Bull. 17 (1994) 306.
[12] F. Yoshida, J.G. Topliss, J. Med. Chem. 43 (2000) 2575–2585.
[13] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, J. Med. Chem. 45 (2002) 2615–2623.
[14] J.J. Lu, K. Crimin, J.T. Goodwin, P. Crivori, C. Orrenius, L. Xing, P.J. Tandler, T.J. Vidmar, B.M. Amore, A.G.E. Wilson, P.F.W. Stouten, P.S. Burton, J. Med. Chem. 47 (2004) 6104–6107.
[15] T.J. Hou, J.M. Wang, W. Zhang, X.J. Xu, J. Chem. Inf. Model. 47 (2007) 460–463.
[16] ADRIANA.Code, version 1.0, Molecular Networks GmbH, Erlangen, Germany. <http://www.molecular-networks.com> (accessed Dec 2007).

[17] Accelrys Inc., Cerius^2 Modelling Environment Release 4.7, Accelrys Inc., San Diego, 2003, <http://www.accelrys.com> (accessed Oct 2006).
[18] <http://chemfinder.cambridgesoft.com/>.
[19] <http://sis.nlm.nih.gov/chemical.html/>.
[20] T.J. Hou, J.M. Wang, W. Zhang, X.J. Xu, J. Chem. Inf. Model. 47 (2007) 208–219.
[21] P. Ertl, B. Rohde, P. Selzer, J. Med. Chem. 43 (2000) 3714–3717.
[22] K.J. Miller, J. Am. Chem. Soc. 112 (1990) 8533–8542.
[23] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, J. Chem. Inf. Comput. Sci. 36 (1996) 1205–1213.
[24] G. Moreau, P. Broto, Nouv. J. Chim. 4 (1980) 359–360.
[25] M. Wagener, J. Sadowski, J. Gasteiger, J. Am. Chem. Soc. 117 (1995) 7769–7775.
[26] J. Gasteiger, M. Marsili, Tetrahedron Lett. 34 (1978) 3181–3184.
[27] J. Gasteiger, M. Marsili, Tetrahedron Lett. 36 (1980) 3219–3228.
[28] J. Gasteiger, M.G. Hutchings, J. Am. Chem. Soc. 106 (1984) 6489–6495.
[29] R.H. Rohrbaugh, P.C. Jurs, Anal. Chim. Acta 199 (1987) 99–109.
[30] D.T. Stanton, P.C. Jurs, Anal. Chem. 62 (1990) 2323–2329.
[31] L. Terfloth, J. Gasteiger, Screening – Trends in Drug Discov. 2 (2001) 49–51.
[32] SONNIA can be obtained from Molecular Networks GmbH, Erlangen, Germany. <http://www.molecular-networks.com> (accessed Dec 2007).
[33] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, second ed. Wiley-VCH, Weinheim, 1999.