Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method

Chang-Ying Ma, Sheng-Yong Yang*, Hui Zhang, Ming-Li Xiang, Qi Huang, Yu-Quan Wei

State Key Laboratory of Biotherapy, West China Hospital, West China Medical School, Sichuan University, Chengdu, Sichuan 610041, PR China

ARTICLE INFO

Article history:
Received 3 December 2007
Received in revised form 10 March 2008
Accepted 19 March 2008
Available online 28 March 2008

Keywords:
Support vector machine
Genetic algorithm
Conjugate gradient method
Plasma protein binding rate
Oral bioavailability

ABSTRACT

In this study, support vector machine (SVM) method combined with genetic algorithm (GA) for feature selection and conjugate gradient (CG) method for parameter optimization (GA–CG–SVM), has been employed to develop prediction models of human plasma protein binding rate (PPBR) and oral bioavailability (BIO). The advantage of the GA–CG–SVM is that it can deal with feature selection and SVM parameter optimization simultaneously. Five-fold cross-validation as well as independent test set method were used to validate the prediction models. For the PPBR, a total of 692 compounds were used to train and test the prediction model. The prediction accuracy by means of 5-fold cross-validation is 86% and that for the independent test set (161 compounds) is 81%. These accuracies are markedly higher over that of the best model currently available in literature. The number of descriptors selected is 29. For the BIO, the training set is composed of 690 compounds and external 76 compounds form an independent validation set. The prediction accuracy for the training set by using 5-fold cross-validation and that for the independent test set are 80% and 86%, respectively, which are better than or comparable to those of other classification models in literature. The number of descriptors selected is 25. For both the PPBR and BIO, the descriptors selected by GA–CG method cover a large range of molecular properties which imply that the PPBR and BIO of a drug might be affected by many complicated factors.

1. Introduction

Currently *in silico* prediction of pharmacokinetic and pharmacodynamic properties, including absorption, distribution, metabolism, excretion and toxicity (ADMET), in the early stage of drug discovery has been thought to be an efficient way to help to reduce the cost of drug development [1,2]. Up to now, many prediction models of ADMET properties have been established [3,4], but some of these models are still not competent for the prediction purpose [5], such as, typically, oral bioavailability (BIO) [6–8] and human plasma protein binding rate (PPBR) [8].

Oral bioavailability is defined as (taking the FDA's definition) “the rate and extent to which the active ingredient or active moiety is absorbed from a drug product and becomes available at the site of action” [9]. The oral bioavailability can affect the dose regimen and determines how much a drug should be given a time [1]. When the drug enters the blood circulation, some of it will bind in plasma to constituent proteins, such as albumin (primarily acidic drugs), α1-Acid glycoprotein (basic drugs), and lipoproteins (neutral and basic drugs) [10]. The human plasma protein binding rate is thus expressed as the percentage of a drug bound to plasma proteins. The BIO and PPBR are generally thought to be affected by many factors. For example, for a drug to be orally bioavailable, it must get to the general circulation by not only passing through the intestine, but also through the liver where it is subject to first-pass metabolism (hepatic clearance) [11]. This implies that many factors, such as its solubility, lipophilicity, gastrointestinal transit and first-pass metabolism may have important impacts on the oral bioavailability for a drug when orally administrated [1,12]. This is probably one of the main reasons why they are difficult to be predicted. On the other hand, most of the prediction models developed so far are based on the quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) [7,13–16]. However, there is an inherent deficiency for the QSAR and QSPR, namely, the limitation of the diverse range of chemical structures [17,18], which may be another reason why the BIO and PPBR are difficult to be predicted. An alternative way to overcome this problem is the use of nonlinear supervised learning method, such as support vector machine (SVM) [8,12,19], which can cover more diverse range of structures than those described by the QSAR and QSPR models.

Although several previous studies have shown that the SVM is capable to correctly predict ADMET-related properties [4,12], there still exist two problems for the SVM, namely feature selec-

* Corresponding author. Tel.: +86 28 85164063; fax: +86 28 85164060.
E-mail address: yangsy@scu.edu.cn (S.-Y. Yang).

0731-7085/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2008.03.023

tion and SVM parameter optimization. The two problems have been shown to be crucial to the prediction efficiency and accuracy of SVM classification [20–22]. In particular, the feature subset selection and optimal SVM parameters setting influence each other, which indicates that they should be dealt with simultaneously [23,24]. Thus, in this investigation, an integrated scheme, in which the feature selection and SVM parameter optimization are considered at the same time, will be used to overcome this type of problem. In this integrated scheme, genetic algorithm (GA) is used for the feature selection and conjugate gradient (CG) method is used for the parameter optimization.

Taking together, in this study, support vector machine method combined with genetic algorithm for feature selection and conjugate gradient method for parameter optimization (GA–CG–SVM), will be used to develop prediction models of PPBR and BIO. The generated SVM prediction models will be validated by 5-fold cross-validation and independent test set method. We organize this paper as follows: The second part presents a detailed description of the proposed integrated GA–CG–SVM scheme. In the third part, we shall apply the integrated GA–CG–SVM method to build SVM prediction models of PPBR and BIO. Validations of the generated models will also be included in this part. Conclusions are offered in the final part.

2. Materials and methods

2.1. Support vector machine

The SVM theory has been extensively described in many literatures [25,26]. Here we just make a short summary to the basic idea of SVM.

In SVM, each object is described by a vector x_i of N real numbers (features or descriptors), which corresponds to a point in an N-dimensional space. The objects in the first class (positive) are each assigned a value of $y_i = +1$, the other ones in the second class (negative) are $y_i = -1$. In linearly separable cases, the objects can be correctly classified by

$$w \cdot x_i + b \geq +1, \quad \text{for} \quad y_i = +1\,(\text{positive}) \tag{1}$$

$$w \cdot x_i + b \leq -1, \quad \text{for} \quad y_i = -1\,(\text{negative}) \tag{2}$$

where w is a vector normal to the hyperplane, b is a scalar quantity. With the w and b being solved, a classifying determination function is obtained as follows:

$$f(x) = \text{sign}[(w \cdot x_i) + b] \tag{3}$$

For a linear non-separable case, no hyperplane can be used to perfectly separate two sets. In this case, we can introduce a nonnegative slack variable $\xi_i \geq 0$, $i = 1, \ldots, m$. Such that

$$w \cdot x_i + b \geq +1 - \xi_i, \quad \text{for} \quad y_i = +1 \tag{4}$$

$$w \cdot x_i + b \leq -1 + \zeta_i, \quad \text{for} \quad y_i = -1 \tag{5}$$

The purpose here is to find a hyperplane that provides the minimum number of training errors. The equation to be solved becomes:

$$\underset{w,b}{\text{Max}} \frac{2}{||w||} + C\sum_{i=1}^{m} \xi_i \quad \text{Subject to} \quad y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0 \tag{6}$$

where C is the penalty parameter, which should be predetermined by user. The parameter C has important impact on the accuracy of the SVM classifier, thus it should be chosen carefully. Similar to the linearly separable cases, Eq. (6) can also be solved by Lagrangian multipliers method.

The nonlinear (non-)separable cases could be easily transferred to linear cases through projecting the input features into a new high-dimensional feature space by using a kernel function $K(x_i, x_j)$. Several types of kernel functions can be used. Due to that the radial basis function (RBF) (see Eq. (7)) performs quite well in many cases, RBF will also be used in this study.

$$k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \tag{7}$$

where γ is a parameter which should be specified by user in advance.

Obviously, to solve a realistic problem using SVM, not only the penalty parameter C, but also the kernel parameter (such as γ in the case of RBF) should be optimized to obtain a better classification model [20,27].

2.2. Parameter optimization by using conjugate gradient method

As stated above, two parameters, namely the penalty parameter C and kernel parameter γ, should be predetermined and optimized to obtain a good model. Traditionally, the method used for optimization of (C, γ) is Grid-search algorithm [28], which is time-consuming. Thus in this investigation, we shall use conjugate gradient method to optimize the (C, γ).

As it has been known, different pairs of (C, γ) give different accuracy. This can be expressed as a function:

$$A = f(C, \gamma) \tag{8}$$

where A is the accuracy of the SVM. Optimizing C and γ means finding an optimal pair of (C, γ) to maximize A. The accuracy of n-fold cross-validation is used to represent the accuracy of SVM since it has been shown that n-fold cross-validation can prevent the overfitting problem [28]. Here 5-fold cross-validation is used due to that 5-fold cross-validation generally performs quite well for middle size problems like those studied here [28,29].

Eq. (8) can be easily transferred to Eq. (9).

$$A' = -f(C', \gamma') \tag{9}$$

where C' is the logarithm of C, γ' is the logarithm of γ, A' is minus accuracy. Now the problem becomes that finding an optimal (C', γ') to minimize A', which can be easily solved by using standard conjugate gradient method.

2.3. Genetic algorithm for the feature selection

Genetic algorithm is a very popular optimization algorithm, which is based on a direct analogy to Darwinian evolutionary ideas of natural selection and genetics in biological systems [23]. GA has been successfully applied to a range of diverse problems such as data mining and optimization. Recently it has also been used for the feature selection in SVM modeling [20,30].

In our implementation of GA, a binary string with each bit representing a feature (0: not selected, 1: selected) was used to represent a chromosome. The GA is then applied to a population of randomly generated binary strings. The fitness of each string is determined as follows:

$$\text{fitness} = W_{\text{A}} \times \text{SVM\_accuracy} + W_{\text{F}} \times N_{\text{F}} \tag{11}$$

where W_{A} is the SVM classification accuracy weight, N_{F} is the number of feature selected, W_{F} is the weight of feature number. The accuracy of 5-fold cross validation was used as SVM_accuracy. W_{A} and W_{F} can be adjusted based on their relative importance.

Roulette wheel selection algorithm was used to choose the chromosomes for crossover to produce offspring. The swapping positions were randomly created and the crossover rate can be adjusted. The mutation was allowed and its rate can be adjusted.

2.4. Experimental data sets for BIO and PPBR

Experimental data sets used in this study were collected from literature. Each compound in these data sets has been carefully checked in order to eliminate the duplication, for example, the same compound but different names are assigned. The compounds were then classified into two categories, positive (+) and negative (−), according to their relative values measured by experiment.

For the human oral bioavailability, a total of 766 compounds were collected [6]. Six hundred and ninety compounds were used to train and test the prediction model and the remaining 76 compounds compose an independent validation set for further assessing the prediction model. These drugs were categorized into two classes: “positive” if its bioavailability $\geq$ 20%, otherwise “negative”. A set of 853 compounds with their human plasma protein binding rate measured experimentally were taken from literature [15]. These compounds were also divided into training set (692 compounds) and the independent validation set (161 compounds). The actual class of each compound was identified based on its relative value of plasma protein binding rate: positive assigned if its plasma protein binding rate $\geq$ 75%, otherwise negative assigned.

2.5. Modeling details

All the structures of the compounds were generated and then optimized by using Cerius2 program package [31]. The 3D structure of each compound was manually inspected to ensure that each molecule was properly represented.

Initially, 951 molecular descriptors were chosen. These descriptors cover 13 different categories (Table 1), including constitutional (48), topological (119), randic molecular profiles (41), geometrical (74), RDF (150), atom-centred fragments (120), walk and path counts (47), connectivity indices (33), edge adjacency indices (107), eigenvalue_based indices (44), functional group counts (154), molecular properties (11), ET-state properties (3).

Table 1
The categories of molecular descriptors initially used in this work

Category of descriptor	Number
Constitutional descriptors	48
Topological descriptors	119
Randic molecular profiles	41
Geometrical descriptors	74
RDF descriptors	150
Atom-centred fragments	120
Walk and path counts	47
Connectivity indices	33
Edge adjacency indices	107
Eigenvalue_based indices	44
Functional group counts	154
Molecular properties	11
ET-state properties	3

Firstly the initial feature subset was preprocessed which purpose is to eliminate the obvious “bad” descriptors and reduce the redundancy and overlapping of the descriptors. In this account, the following descriptors are removed: (1) descriptors with too many zero values, (2) descriptors with very small standard deviation values (<0.5%), and (3) descriptors which are highly correlated with others (correlation coefficients >90%). After the preprocessing, the descriptor values were scaled to a range of -1 to $+1$, which is necessary since the different ranges of descriptor values will influence the quality of the SVM modeling.

The termination criteria for the running of genetic algorithm are that the generation number reaches generation 200 or that the fitness value does not improve during the last 10 generations. The crossover rate was set to 0.8 and mutation rate 0.05.

Performance of the SVM classification model can be assessed by the quantity of true positives (TP, true PPBR+/BIO+ agents), true negatives (TN, true PPBR−/BIO− agents), false positives (FP, false PPBR+/BIO+ agents), false negatives (FN, false PPBR−/BIO− agents). Sensitivity $SE = TP/(TP + FN)$ and specificity $SP = TN/(TN + FP)$ are the prediction accuracies for the PPBR+/BIO+ and PPBR−/BIO−,

Table 2
Descriptors selected by the automatic feature selection process in the SVM modeling for human plasma protein binding rate

Name	Description	Class
Mv	Mean atomic van der Waals volume (scaled on carbon atom)	Constitutional descriptors
MS	Mean electrotopological state	Constitutional descriptors
nAT	Number of atoms	Constitutional descriptors
nS	Number of sulfur atoms	Constitutional descriptors
ARR	Aromatic ratio	Constitutional descriptors
PJI2	2D Petitjean shape index	Topological descriptors
PW2	Path/walk 2-Randic shape index	Topological descriptors
ZM2v	Second Zagreb index by valence vertex degrees	Topological descriptors
SRW05	Self-returning walk count of order 05	Walk and path counts
X4Av	Average valence connectivity index chi-4	Connectivity indices
J3D	3D-Balaban index	Geometrical descriptors
RDF020m	Radial distribution function – 2.0/weighted by atomic masses	RDF descriptors
RDF035m	Radial distribution function – 3.5/weighted by atomic masses	RDF descriptors
RDF075m	Radial distribution function – 7.5/weighted by atomic masses	RDF descriptors
C-001	CH_3R/CH_4	Atom-centred fragments
C-003	CHR_3	Atom-centred fragments
C-011	CR_3X	Atom-centred fragments
C-017	$=CR_2$	Atom-centred fragments
H-047	H attached to $C^1(sp^3)/C^0(sp^2)$	Atom-centred fragments
H-048	H attached to $C^2(sp^3)/C^1(sp^2)/C^0(sp)$	Atom-centred fragments
H-049	H attached to $C^3(sp^3)/C^2(sp^2)/C^3(sp^2)/C^3(sp)$	Atom-centred fragments
H-052	H attached to $C^0(sp^3)$ with 1X attached to next C	Atom-centred fragments
S-107	$R_2S/RS-SR$	Atom-centred fragments
EEig03x	Eigenvalue 03 from edge adj. matrix weighted by edge degrees	Edge adjacency indices
ESpm01d	Spectral moment 01 from edge adj. matrix weighted by dipole moments	Edge adjacency indices
ALOGP2	Squared Ghose–Crippen octanol–water partition coefficient ($\log P^2$)	Molecular properties
TPSA(Tot)	Topological polar surface area using N, O, S, P polar contributions	Molecular properties
nCconj	Number of non-aromatic conjugated $C(sp^2)$	Functional group counts
nRCONHR	Number of secondary amides (aliphatic)	Functional group counts

Table 3
Prediction accuracies of the SVM models of PPBR and BIO by using 5-fold cross-validation

Activity	Cross-validation	Positive			Negative			Q (%)
		TP	FN	SE (%)	TN	FP	SP (%)	
PPBR	1	68	8	89	52	10	84	87
	2	65	11	86	54	8	87	86
	3	67	9	88	48	14	77	83
	4	67	9	88	52	10	84	86
	5	67	10	87	58	5	92	89
	Average			88			85	86
BIO	1	101	2	98	7	27	21	79
	2	103	0	100	5	30	14	78
	3	99	4	96	10	25	29	79
	4	103	0	100	10	25	29	82
	5	103	1	99	11	24	31	82
	Average			99			25	80

respectively. The overall accuracy (Q) is calculated by the equation: $Q = (TP + TN)/(TP + TN + FP + FN)$.

All the calculations were carried out by the in-house GA–CG–SVM program, which calls the gcc package of libsvm (version 2.83) for the SVM calculation [32]. Molecular descriptors were generated using the online program PCLIENT [33].

3. Results and discussion

3.1. Prediction model for human plasma protein binding rate

The training set containing 692 compounds was used to train the SVM prediction model of PPBR. The number of descriptors initially chosen is 951. Passing through feature preprocessing and automatic feature selection processes, the number of descriptors finally selected for building the SVM model is 29. Table 2 lists the selected descriptors and their descriptions. These molecular descriptors fall into several categories as follows: constitutional descriptors (5), topological descriptors (3), Walk and path counts (1), connectivity indices (1), geometrical descriptors (1), RDF descriptors (3), atom-centred fragments (9), edge adjacency indices (2), molecular properties (2), functional group counts (2). Obviously the descriptors represent different types of molecular properties, which at least to some extent reflect that the PPBR of a drug is affected by many complicated factors.

The SVM prediction model generated was tested by 5-fold cross-validation method. The prediction accuracies for the training set by using 5-fold cross-validation are shown in Table 3. The average prediction accuracies for the positive (SE) and negative (SP) are 88% and 85%, respectively. The average overall prediction accuracy is 86%.

The purpose of the SVM model generated is not just to classify the training set agents correctly into PPBR+ and PPBR−, but also to verify whether the SVM model is capable of classifying external agents of the validation set series accurately as PPBR+ or PPBR−. Thus an independent validation set comprising 161 compounds was further used to evaluate the model just built. For the independent validation set, the prediction accuracies for positive, negative and overall are 72%, 89% and 81% (Table 4).

Trotter and Holden [8] developed several prediction models for PPBR by using different supervised machine learning methods, including SVM, ANN, C5.0, and neighbours. The training set they used just contains 240 compounds, which is much smaller than that we used. The overall prediction accuracies are between 69% and 74% (see Table 4). Apparently our GA–CG–SVM model for PPBR is superior over those developed by other methods.

3.2. Prediction model for oral bioavailability

Six hundred and ninety compounds composing the training set were used to train and test the SVM prediction model of BIO. Again the number of initial descriptors was 951. And it was reduced to 25 after preprocessing and automatic feature selection processes. Table 5 presents the descriptors as well as their definitions. Similar to PPBR, the descriptors include different types of molecular properties. Again, this implies that oral bioavailability of a drug might be affected by many factors.

The prediction accuracies for the training set by using 5-fold cross-validation are shown in Table 3. Apparently acceptable overall prediction accuracy (average: 80%) is obtained. The average prediction accuracy for the positive reaches 99%. But the average accuracy for negative is just 25%, which indicates that the negative could not be identified correctly. This is nothing to be surprised since very few of prediction accuracies for the negative are bigger than 50% among all of the classification models reported so far (also see the next paragraph). Further, the independent validation set composed of 76 compounds was used to evaluate the performance of the model just created. The overall prediction accuracy for the independent validation set is 86% with an accuracy of 97% for positive and 44% for negative (Table 6).

Trotter and Holden [8] built several prediction models by using different SVM methods, ANN, RBF, C5.0 and neighbours, in which the training set they used contains 240 compounds and test set 241 compounds. The highest overall prediction accuracy is 87% (Table 6) which model was derived by RBF SVM, another implementation of SVM. Notably, the prediction accuracies for the positive are around 90%, but those for the negative are less than 50% (Table 6), which are very similar with our results. Wang et al. [12] used a smaller training set (133 compounds) and test set (34 compounds) to build two prediction models by using the linear discriminant analysis (LDA) and grid-search based support vector machine (GS-SVM). The GS-SVM gives the highest overall prediction accuracy (86% for training

Table 4
Prediction accuracies of different models of PPBR by independent validation set or test set

Method	Data set	No. of descriptors	Q (%)	Positive			Negative		
				TP	FN	SE (%)	TN	FP	SP (%)
GA–CG–SVM	Validation (161)	29	81	53	21	72	77	10	89
GA–SVM	Validation (161)	104	66	67	7	91	39	48	45
Lin. SVM [8]	Test (241)	16	72			73			69
Quad. SVM [8]	Test (241)	16	72			87			57
RBF SVM [8]	Test (241)	16	71			85			48
ANN [8]	Test (241)	16	72			78			61
RBF [8]	Test (241)	16	74			84			54
C5.0 [8]	Test (241)	16	71			84			45
Neighbours [8]	Test (241)	16	69			78			53

Table 5
Descriptors selected by the automatic feature selection process in the SVM modeling for BIO

Name	Description	Class
MW	Molecular weight	Constitutional descriptors
Ms	Mean electrotopological state	Constitutional descriptors
nR05	Number of 5-membered rings	Constitutional descriptors
S2K	2-Path Kier alpha-modified shape index	Topological descriptors
ZM1V	First Zagreb index by valence vertex degrees	Topological descriptors
BLI	Kier benzene-likeliness index	Topological descriptors
HNar	Narumi harmonic topological index	Topological descriptors
TI2	Second Mohar index TI2	Topological descriptors
D/Dr11	Distance/detour ring index of order 11	Topological descriptors
X1A	Average connectivity index chi-1	Connectivity descriptors
X0Av	Average valence connectivity index chi-0	Connectivity descriptors
X3v	Valence connectivity index chi-3	Connectivity descriptors
MWC02	Molecular walk count of order 02	Walk and path counts
ASP	Asphericity	Geometrical descriptors
QYYm	Qyy COMMA2 value/weighted by atomic masses	Geometrical descriptors
DISPe	d COMMA2 value/weighted by atomic Sanderson electronegativities	Geometrical descriptors
RDF035m	Radial distribution function – 3.5/weighted by atomic masses	RDF descriptors
RDF040m	Radial distribution function – 4.0/weighted by atomic masses	RDF descriptors
C-005	CH_3X	Atom-centred fragments
H-048	H attached to $C^2(sp^3)/C^1(sp^2)/C^0(sp)$	Atom-centred fragments
H-052	H attached to $C^0(sp^3)$ with 1X attached to next C	Atom-centred fragments
O-060	Al—O—Ar/Ar—O—Ar/R—O—R/R—O—C=X	Atom-centred fragments
ESpm01d	Spectral moment 01 from edge adj. matrix weighted by dipole moments	Edge adjacency indices
MLOGP	Moriguchi octanol–water partition coefficient	Molecular properties
nCconj	Number of non-aromatic conjugated $C(sp^2)$	Functional group counts

Table 6
Prediction accuracies of different models of BIO by independent validation set or test set

Method	Data sets	No. of descriptors	Q (%)	Positive			Negative		
				TP	FN	SE (%)	TN	FP	SP (%)
GA–CG–SVM	Validation (76)	25	86	58	2	97	7	9	44
GA–SVM	Validation (76)	11	79	58	2	97	2	14	20
Lin. SVM [8]	Test (241)	68	79			85			46
Quad. SVM [8]	Test (241)	68	83			89			39
RBF SVM [8]	Test (241)	68	87			95			36
ANN [8]	Test (241)	68	85			92			43
RBF [8]	Test (241)	68	86			95			27
C5.0 [8]	Test (241)	68	83			90			39
Neighbours [8]	Test (241)	68	83			90			36
LDA [13]	Test (34)	5	82						
GS-SVM [13]	Test (34)	5	85						

set and 85% for test set). Compared with these previous models, the training set we used is much larger (690 compounds), which is the largest dataset for BIO so far as far as we know. The prediction accuracies of our GA–CG–SVM model are not better but comparable to those of the previous models. The fact that the prediction accuracy of BIO is difficult to be improved further implies that the bioavailability of a drug is typically affected by many complicated factors.

3.3. Influence of the use of parameter optimization

As stated in Section 1, the feature subset selection and optimal SVM parameters setting are two key factors to influence the efficiency and accuracy of SVM classification model. In fact, the effect of the feature selection on the SVM model has been widely discussed in literature [18,34,35]. But few of studies have analyzed the possible impact of the SVM parameter optimization [20]. In the follows, we shall use the same data sets as before to rebuild SVM models in which just feature selection (by GA) was performed. The purpose here is to demonstrate the possible influence of the use of SVM parameter optimization.

The prediction accuracies of the rebuilt models by using GA–SVM are shown in Tables 4 and 6 for PPBR and BIO, respectively. For the PPBR, the overall accuracies are 68% (obtained by 5-fold cross-validation) and 66% for the training set and independent validation set, respectively, which are lower by 18% and 15% compared with the corresponding values of GA–CG–SVM models, respectively. For the individual categories, the prediction accuracies of GA–SVM models are also lower than those of GA–CG–SVM models (see Table 4). For the BIO, the prediction accuracies of GA–SVM for the negative and the global are obviously lower than the corresponding values of GA–CG–SVM model (see Table 6). These results clearly show that the parameter optimization is very important for improving the prediction accuracy of SVM models.

4. Conclusions

In this account, SVM method combined with GA for feature selection and CG method for parameter optimization (GA–CG–SVM), has been employed to develop classification models of human plasma protein binding rate and human oral bioavailability. The advantage of the GA–CG–SVM is that it can deal with feature selection and SVM parameter optimization simultaneously.

The generated models were validated by using the 5-fold cross-validation and independent validation set methods. For the PPBR, the training set contains 692 compounds and the test set includes external 161 compounds. The prediction accuracy for the training set by using 5-fold cross-validation is 86%. And the prediction accuracy for the independent validation set is 81%. The prediction accuracies obtained here are much higher over that of the best model currently available in literature. The number of descriptors selected by GA–CG is 29. For the BIO, the training set is composed of 690 compounds and the independent validation set contains 76 compounds. The overall prediction accuracies are 80% and 86% for the training set and the independent validation set, respectively, which are better than or comparable to those of other classification models. The number of descriptors selected is 25. For both the PPBR and BIO, the descriptors selected by GA–CG method cover a large range of molecular properties which imply that the PPBR and BIO of a drug are affected by many complicated factors. Finally the influence of the use of parameter optimization on the SVM prediction model has also been examined. The prediction accuracies of the models with parameter optimization involved in the modeling are much higher over those without parameter optimization involved. These results clearly show that the parameter optimization is very important for improving the prediction accuracy of SVM models.

Acknowledgements

This work was supported by the 863 Hi-Tech Program (2006AA020402), the National Natural Science Foundation of China (30772651) and the Youth Foundation of Sichuan Province.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jpba.2008.03.023.

References

[1] H. van de Waterbeemd, E. Gifford, Nat. Rev. Drug Discov. 2 (2003) 192–204.
[2] P.A. Smith, M.J. Sorich, L.S.C. Low, R.A. McKinnon, J.O. Miners, J. Mol. Graphics Modell. 39 (2005) 91–103.
[3] Y.H. Zhao, M.H. Abraham, A. Ibrahim, P.V. Fish, S. Cole, M.L. Lewis, M.J. de Groot, D.P. Reynolds, J. Chem. Inf. Model 47 (2007) 170–175.
[4] J. Huang, G. Ma, I. Muhammad, Y. Cheng, J. Chem. Inf. Model 47 (2007) 1638–1647.
[5] T.I. Opera, H. Matter, Curr. Opin. Chem. Biol. 8 (2004) 349–358.
[6] T. Hou, J. Wang, W. Zhang, X. Xu, J. Chem. Inf. Model 47 (2007) 460–463.
[7] C.W. Andrews, L. Bennett, L.X. Yu, Pharm. Res. 17 (2000) 639–644.
[8] M.W.B. Trotter, S.B. Holden, QSAR Comb. Sci. 22 (2003) 533–548.
[9] M.L. Chen, V. Shah, R. Patniak, W. Adams, A. Hussain, D. Conner, M. Mehta, H. Malinowski, J. Lazor, S.M. Huang, D. Hare, L. Lesko, D. Sporn, R. Williams, Pharm. Res. 18 (2001) 1645–1648.
[10] D.E. Mager, Adv. Drug Deliv. Rev. 58 (2006) 1326–1356.
[11] H. van de Waterbeemd, H. Lennernäs, P. Artursson, Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability, Wiley-VCH, Weinheim, Germany, 2003.
[12] J. Wang, H. Du, X. Yao, Z. Hu, Anal. Chim. Acta 601 (2007) 156–163.
[13] J.V. Turner, D.J. Maddalena, S. Agatonovic-Kustrin, Pharm. Res. 21 (2004) 68–82.
[14] J. Wang, G. Krudy, X. Xie, C. Wu, G. Holland, J. Chem. Inf. Model 46 (2006) 2674–2683.
[15] J.R. Votano, M. Parham, L. Mark Hall, L.H. Hall, L.B. Kier, S. Oloff, A. Tropsha, J. Med. Chem. 49 (2006) 7169–7181.
[16] M. Paul Gleeson, J. Med. Chem. 50 (2007) 101–112.
[17] A.M. Davis, R.J. Riley, Curr. Opin. Chem. Biol. 8 (2004) 378–386.
[18] H. Li, C.W. Yap, Y. Xue, Z.R. Li, C.Y. Ung, L.Y. Han, Y.Z. Chen, Drug Dev. Res. 66 (2006) 245–259.
[19] F. Yoshida, J.G. Topliss, J. Med. Chem. 43 (2000) 2575–2585.
[20] C.L. Huang, C.J. Wang, Expert Syst. Appl. 31 (2006) 231–240.
[21] H. Fröhlich, O Chapelle, Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, Sacramento, CA, USA, 2003.
[22] H. Fröhlich, J. Wegener, A. Zell, QSAR Comb. Sci. 23 (2004) 311–318.
[23] L. Davis, Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.
[24] B. Schølkopf, J. Platt, T. Hofmann (Eds.), Advances in Neural Information Processing Systems (NIPS 2006), Vancouver, Canada, 2007.
[25] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[26] C.J.C. Burges, Data Min. Knowl. Disc. 2 (1998) 127–167.
[27] C.-H. Wu, G.-H. Tzeng, Y.-J. Goo, W.-C. Fang, Expert Syst. Appl. 32 (2007) 397–408.
[28] C.W. Hsu, C.C. Chang, C.J. Lin, A Practical Guide to Support Vector Classification, 2003. Available at: http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf.
[29] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, J. Chem. Inf. Comput. Sci. 44 (2004) 1630–1638.
[30] J. Yang, V. Honavar, IEEE Intell. Syst. 13 (1998) 44–49.
[31] Cerius2, Version 4.11, http://www.accelrys.com.
[32] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. Available at: http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
[33] (a) I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, J. Comput. Aid. Mol. Des. 19 (2005) 453–463;
(b) VCCLAB, Virtual Computational Chemistry Laboratory, 2005. http://www.vcclab.org.
[34] H. Yu, J. Yang, W. Wang, J. Han, in: F. Fitsworth (Ed.), Proceedings of the 2003 IEEE Bioinformatics Conference, United States of America by Victor Graphics, Inc, Canifornia, USA, 2003, pp. 220–228.
[35] S. Degroeve, B. De Baets, Y. Van de Peer, P. Rouze, Bioinformatics 18 (2002) s75–s83.