Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 14 (2006) 4834–4841

Bioorganic & Medicinal Chemistry

QSAR study of 1,4-dihydropyridine calcium channel antagonists based on gene expression programming

Hong Zong Si,[a,b,*] Tao Wang,[c] Ke Jun Zhang,[d] Zhi De Hu[a,*] and Bo Tao Fan[e]

[a] *Department of Chemistry, Lanzhou University, 730000 Lanzhou, PR China*
[b] *Center for Disease Control of Gansu Province, 730020 Lanzhou, PR China*
[c] *Clinical Laboratory, The First Hospital of Lanzhou University, 73000 Lanzhou, PR China*
[d] *School of Mechanical and Electrical Engineering, JUST, 341000 Ganzhou, PR China*
[e] *Université Paris 7-Denis Diderot, ITODYS 1, rue Guy de la Brosse, 75005 Paris, France*

Received 26 January 2006; revised 13 March 2006; accepted 13 March 2006
Available online 31 March 2006

Abstract—The gene expression programming, a novel machine learning algorithm, is used to develop quantitative model as a potential screening mechanism for a series of 1,4-dihydropyridine calcium channel antagonists for the first time. The heuristic method was used to search the descriptor space and select the descriptors responsible for activity. A nonlinear, six-descriptor model based on gene expression programming with mean-square errors 0.19 was set up with a predicted correlation coefficient (R^2) 0.92. This paper provides a new and effective method for drug design and screening.

1. Introduction

Calcium channel antagonists (CCA) have been established as one of the first line drugs for treatment of hypertension because of their promising depressor effect and relatively good tolerability.[1] The 1,4-dihydropyridines (1,4-DHPs), are a class of CCA and used most frequently as antihypertensive and the 1,4-DHPs may lead to other beneficial effects such as regression of left ventricular pressure and vascular hypertrophy, renal protection, weak anti-platelet, antiischemic and antiatherogenic activity.[2–4] The mechanism of these drugs acts directly on the voltage-dependent calcium channels localized in the cell membrane and blocks the flux of calcium ions from the extracellular medium to the cell cytoplasm. In general, IC_{50} (the molar concentration of the drug required to inhibit 50% of the contraction of guinea pig ileum induced by methyl-furmethide) is used to evaluate the efficiency of a drug. However, the optimal interaction and therapeutic efficacy of the compound depend on its chemical structure. Therefore, quantitative structure–property/activity relationship (QSPR/QSAR) was proposed for predicting the IC_{50} of 1,4-DHP.[5–12] The advances in QSAR studies have widened the scope of rationalizing drug design and the search for the mechanisms of drug actions.[13–15] In addition, they are useful in areas such as design of virtual compound libraries, computational-chemical optimization of compounds, and design of combinatorial libraries with appropriate ADME (absorption, distribution, metabolism, and excretion) properties. Method of building the QSAR model plays a key role for the quality of the models. The gene expression programming (GEP) is a novel algorithm developed from the machine learning community at present. Due to its remarkable generalization performance, for the first time it was used for QSAR model building. GEP also based on descriptors was calculated from the molecular structure alone by the software CODESSA. GEP has been successfully used to predict evaporation estimation[16] and cement strength.[17] In order to investigate the influence of different descriptors on IC_{50}, HM was used to build several multivariable linear models. The aim of this study was to explore the IC_{50} of 1,4-DHP with diverse structures and establish a new and accurate QSPR model. The use of quantitative model not only can narrow the search for future drug compounds, but also gain some insight into the structural factors that are responsible for their activities. The prediction results were very satisfactory in regression analyses, which proved GEP to be a powerful and useful tool in drug design and discovery.

Keywords: QSAR; Calcium channel antagonists; Gene expression programming.

[*] Corresponding author. Tel.: +86 931 8912587; fax: +86 931 8912582; e-mail: sihz03@st.lzu.edu.cn

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.03.019

2. Results and discussion

2.1. The results of HM

Six hundred and four descriptors were calculated by the CODESSA program for all the compounds. To select the set of descriptors that are most relevant to the IC_{50} of 1,4-DHP, the linear models with 6 variables were built.

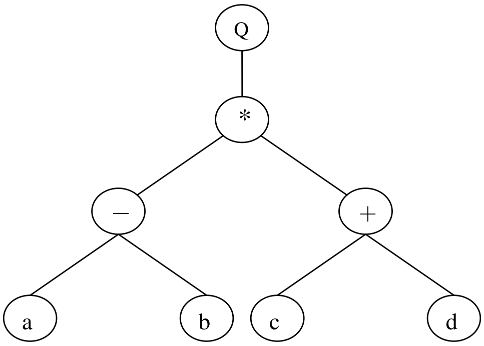


Figure 1. GEP expression tree.

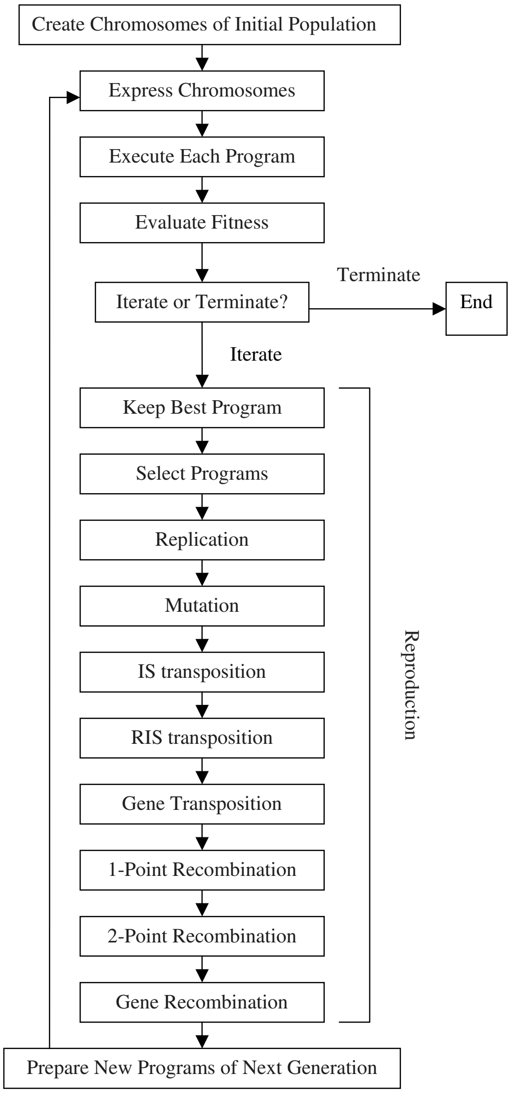
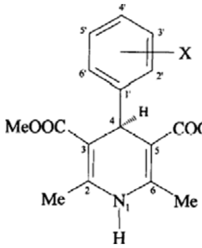


Figure 2. The flowchart of gene expression algorithm.

Table 1. Experimental and predicted activities ($-\log(IC_{50})$) of 1,4-dihydropyridine calcium channel antagonists



Antagonist	X	$\log(1/IC_{50})$		
		Experimental	HM	GEP
1	3′-Br	8.89	8.22	8.76
2[a]	2′-CF_3	8.82	8.74	9.25
3	2′-Cl	8.66	8.02	8.10
4	3′-NO_2	8.40	8.20	8.23
5[a]	2′-CH=CH_2	8.35	8.33	7.96
6	2′-NO_2	8.29	7.95	8.16
7	2′-Me	8.22	7.71	7.67
8[a]	2′-Et	8.19	8.05	8.63
9	2′-Br	8.12	7.45	7.68
10	2′-CN	7.80	8.37	8.64
11[a]	3′-Cl	7.80	7.54	8.31
12	3′-F	7.68	7.78	7.81
13	H	7.68	7.95	7.66
14[a]	3′-CN	7.46	7.78	7.75
15	3′-I	7.38	7.21	7.09
16	2′-F	7.37	7.59	7.63
17[a]	2′-I	7.33	7.19	7.31
18	2′-OMe	7.24	6.63	7.14
19	3′-CF_3	7.13	7.37	7.64
20[a]	3′-Me	6.96	6.70	6.43
21	2′-OEt	6.96	7.19	7.53
22	3′-OMe	6.72	6.58	6.76
23[a]	3′-NMe_2	6.05	5.98	6.06
24	3′-OH	6.00	6.15	6.06
25	3′-NH_2	5.70	5.05	5.60
26[a]	3′-OAc	5.22	5.46	5.49
27	3′-OCOPh	5.20	4.57	5.64
28	2′-NH_2	4.40	5.43	4.86
29[a]	4′-F	6.89	6.62	6.53
30	4′-Br	5.40	5.69	5.74
31	4′-I	4.64	4.53	5.05
32[a]	4′-NO_2	5.50	5.00	5.34
33	4′-NMe_2	4.00	3.97	4.25
34	4′-CN	5.46	5.99	5.85
35[a]	4′-Cl	5.09	6.02	5.79
36	2′,6′-Cl_2	8.72	8.70	8.71
37	F5	8.36	7.78	8.18
38[a]	2′-F, 6′-Cl	8.12	8.07	8.38
39	2′3′-Cl_2	7.72	7.86	7.67
40	2′-Cl, 5′-NO_2	7.52	8.38	7.30
41[a]	3′,5′-Cl_2	7.03	7.92	7.91
42	2′-OH, 5′-NO_2	7.00	7.11	6.84
43	2′,5′-Me_2	7.00	7.13	7.52
44[a]	2′,4′-Cl_2	6.40	6.71	6.61
45	2′,4′,5′-$(OMe)_3$	3.00	3.18	3.36

[a] Test set.

Table 2. Correlation matrix of the 6 descriptors[a]

Descriptors	HOMOE	MIC	XYSR	YZSR	MSA	THCMD
HOMOE	1.00	0.18	0.12	−0.12	0.09	0.12
MIC		1.00	0.34	−0.07	−0.65	−0.26
XYSR			1.00	0.15	−0.27	−0.19
YZSR				1.00	−0.15	−0.15
MSA					1.00	0.14
THCMD						1.00

[a] HOMOE, HOMO energy; MIC, Moment of inertia C; XYSR, XY Shadow/XY Rectangle; YZSR, YZ Shadow/YZ Rectangle; MSA, Molecular surface area; THCMD, Tot hybridization component of the molecular dipole.

$$\begin{aligned}\mathrm{Log}(1/\mathrm{IC}_{50}) = {} & 0.31–2.55\mathrm{HOMOE} + 2.80 \\ & \times 10^{2}\mathrm{MIC} + 7.07\mathrm{XYSR}–1.60 \\ & \times 10^{-1}\mathrm{YZSR}–3.04 \times 10^{-2}\mathrm{MSA} \\ & –1.16\mathrm{THCMD} \\ & R^2 = 0.90,\ R^2_{CV} = 0.86,\ F = 58.18, \\ & s^2 = 0.22. \qquad (1)\end{aligned}$$

The involved molecular descriptors and their corresponding physical-chemical meaning are given in Table 2. Figure 3 shows the plots of R^2 and R^2_{CV} for the training set as a function of the number of descriptors for the 6-parameter model. R^2 and R^2_{CV} increased with increasing the number of descriptors. Comparing with R^2 and R^2_{CV} from 1 to 6-parameter we found that the 6-parameter model is the best model (Fig. 4). So it was chosen as the best linear model and the corresponding descriptors were used as inputs for the nonlinear model.

2.2. Results of GEP

The fitness was evaluated by Eq. 8, choosing $R = 22.22$, for $n = 45$, so making $f_{\max} = 45 * 22.22 = 1000$. Figures 5 and 6 showed the fitting situation of training set and test set.

The software automatic problem solver (APS)[18] was used to model this function because it allows the easy optimization of intermediate solutions and the easy testing of the evolved models against a test set. Good solution, with an R^2 of 0.92 (Table 3), was obtained. With APS we can convert the C^{++} function into equation.

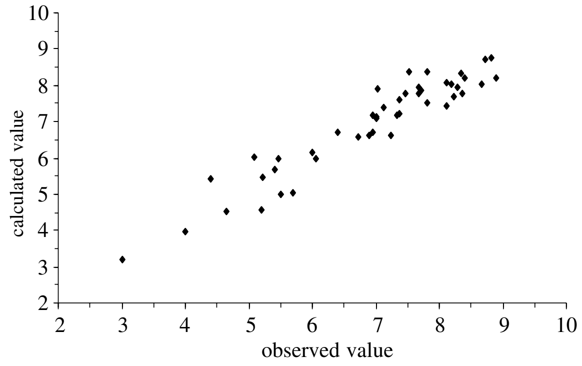


Figure 3. Plot of predicted $\log(1/\mathrm{IC}_{50})$ versus experimental values for the training and testing sets by HM.

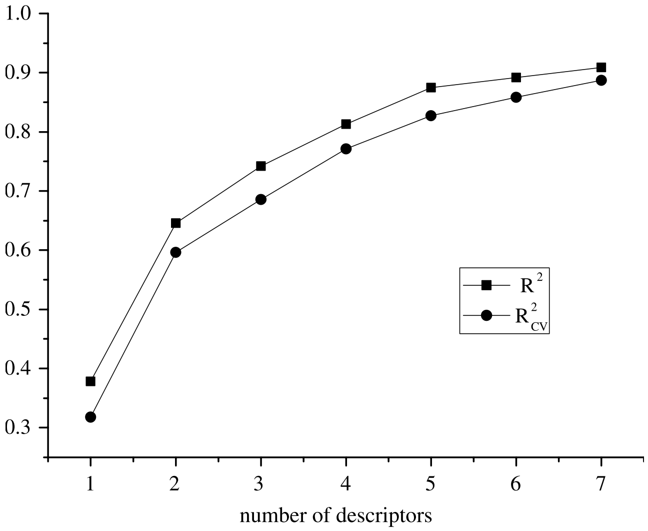


Figure 4. Influence of the number of descriptors on the correlation coefficient (R^2) and the cross-validation correlation coefficient (R^2_{CV}) of the regression model.

Table 3. Result of correlation coefficient (R^2) and mean-square errors (S^2) with GEP and HM

Methods	Training set		Test set	
	R^2	S^2	R^2	S^2
GEP	0.93	0.18	0.88	0.20
HM	0.91	0.21	0.86	0.22

$$\begin{aligned} y = {} & \sin\left(9.76 \times x_3 - x_6 - \frac{x_1}{x_4 - x_2} + x_2\right) \\ & + \sin(\sin(\sin(\sin((\sinh(ceil\sqrt{x_5})))))) \\ & + \sin\left(x_1 - \frac{\cos(x_5)}{f(\frac{x_2}{x_6} - x_3)}\right)\sin(\cos(\sqrt{x_5})) + \mathrm{e}^{x^3} \\ & + p\left(\ln x_{x_5}^{\left(\lg\left(\frac{f(x_4)}{x_6}\right)\right)}\right) + x_2 + p\left(\frac{\sinh(x_2)}{\sin\left(\frac{1}{\sin(x_5)}\right)}\right) \\ & + \sin(\sin(x_1)) \end{aligned}$$

$$f(x) = \ln\left(\frac{\sqrt{-x^2 + 1}}{x}\right) \qquad -1 \leqslant x \leqslant 1$$

$$p(x) = \arctan(x) + \arctan(1) \qquad (2)$$

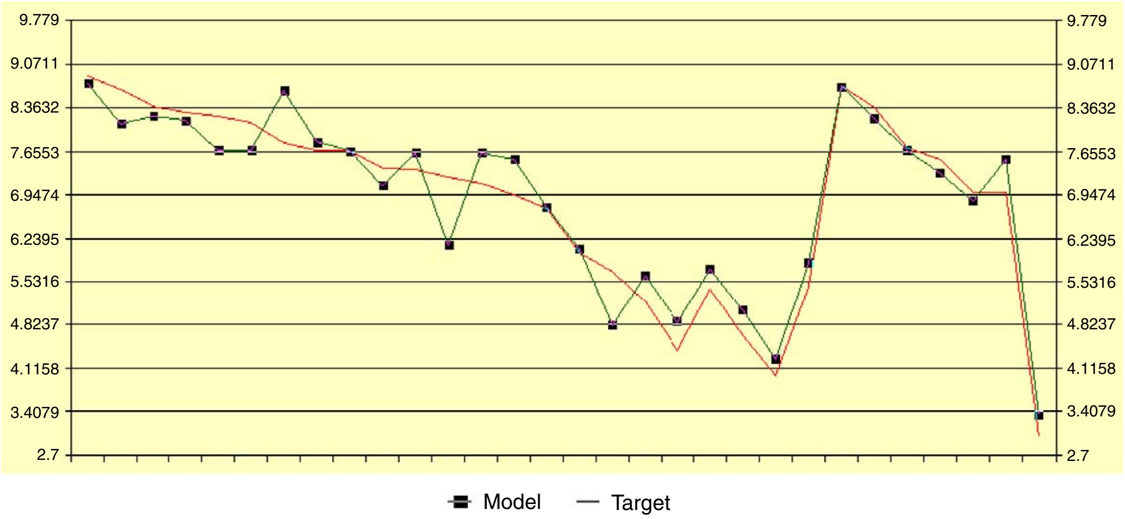


Figure 5. Fitting curve of training set.

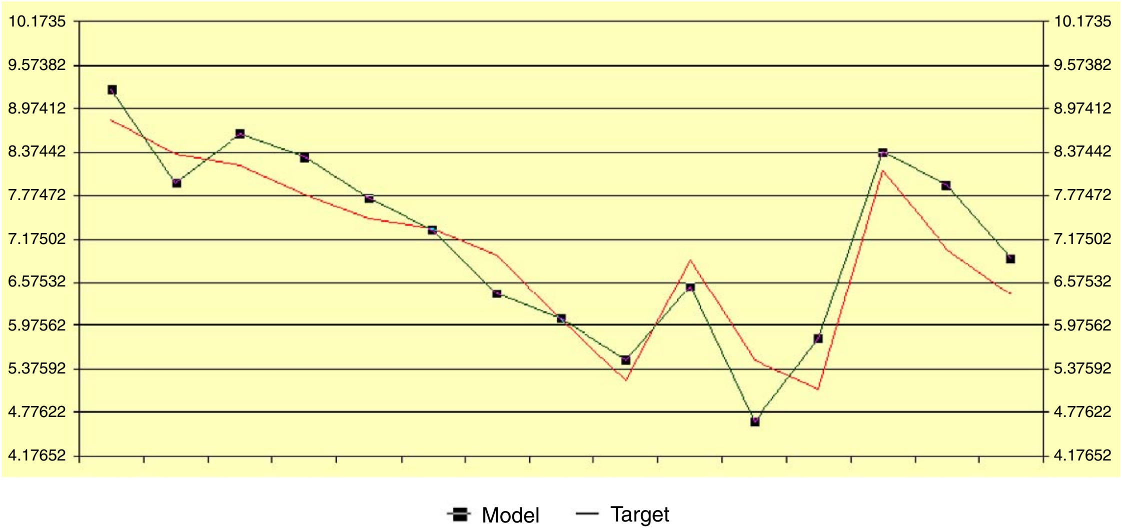


Figure 6. Fitting curve of test set.

We compared the predicted results of GEP and GA-based PLS (GAPLS). Six-descriptor model based on GEP with correlation coefficient (R^2) and mean-square errors 0.92 and 0.19, while the model was based on GAPLS with correlation coefficient (R^2) and mean-square errors 0.80 and 0.43, respectively. Therefore, the predicted ability of GEP is better than that of GAPLS.[24]

2.3. Discussion of the descriptors

By interpretation of the descriptors used, it is possible to gain some insight into factors that are likely to govern the depletion of IC_{50} of 1,4-dihydropyridine. In this work, six descriptors involved in the model belong to two types.

(1) *Moment of inertia C*, *XY Shadow/XY Rectangle*, *YZ Shadow/YZ Rectangle*, and *Molecular surface area* are geometric descriptors. Geometric descriptors encode the structural characteristics related to connectivity, complexity, and shape, proving that the hydrophobic and steric interactions are very important for binding between the antagonists and receptor. *Moments of inertia C* are in the rigid rotator approximation, the principal moments of inertia of a molecule.[19] Positive coefficient of *Moment of inertia C* means that the higher atomic mass and denoting the distance of the *i*th atomic nucleus from the main rotational axes *z* of the molecule will increase $\log(1/IC_{50})$ of 1,4-DHP. By the orientation of the molecule in the space along the axes of inertia the areas of the shadows of the molecule are projected on the XY and YZ planes.[20] The normalized shadow areas are calculated as the ratios $S1/(X_{max}Y_{max})$ and $S2/(Y_{max}Z_{max})$, where X_{max} and Y_{max} are the maximum dimensions of the molecule along the corresponding axes. These values can be obtained by applying 2D square grid on the molecular projection and by summation of the areas of squares overlapped with a projection. Thus, those indices reflect the size (natural

shadow indices) and geometrical shape (normalized shadow indices) of the molecule. However, the coefficient of XYSR (*XY Shadow/XY Rectangle*) is positive, while the coefficient of YZSR (*YZ Shadow/YZ Rectangle*) is negative. Therefore, the orientation of the molecule in the space along the axes of inertia the areas of the shadows may be a key factor which affects the activity of 1,4-DHP. The molecule can be divided into a number of slices along the X axis with the step dx, then for each slice a set of circles corresponding to intersections of the van der Waals spheres and cutting plane is generated. Finally for each circle, the lengths of nonoccluded arcs are found using the step dl, and the surface area is calculated as the sum of the lengths of nonoccluded arcs multiplied by the slice thickness dx.[21] Gaudio et al. found that the van der Waals volume of the entire molecule correlates well with the biological activity. The high negative coefficient of MSA can decrease log(1/IC_{50}) of 1,4-DHP. Larger MSA of DHP has lower log(1/IC_{50}).

(2) *HOMO energy* and *Tot hybridization component of the molecular dipole* are quantum-chemical descriptors. The highest occupied molecular orbital (HOMO) is the highest-energy orbital with one or two electrons.[22,23] *Tot hybridization component of the molecular dipole* represents or depends directly on the quantum-chemically calculated charge distribution in the molecules and therefore describes the polar interactions between molecules or their chemical reactivity. Two quantum-chemical descriptors have negative coefficients. Therefore, log(1/IC_{50}) decreases with the values of HOMOE and THCMD. Analysis of the results obtained indicates that the selected molecular descriptors calculated solely from structures can describe the structural features of the compounds responsible for their biological activity. These two descriptors indicate the importance of the intramolecular electronic effects and reactivity of a molecule in determining the binding ability between the molecule and receptor. Analysis of the results obtained indicates that the selected molecular descriptors calculated solely from structures can describe the structural features of the compounds responsible for their biological activity.

From the results we found that: (1) With the positions of substituents changed from 2′ to 4′, the log(1/IC_{50}) decreases gradually. (2) At the same position, volume of molecule also influenced on the value of log(1/IC_{50}), with the volume increase, the value of log(1/IC_{50}) decreases. (3) The two para-substituted DHPs (H and F) exhibit high activity, while the six para-substituted DHPs (NO_2, CN, Br, Cl, I, and NMe) exhibit low activity. These confirmed the reliability of the model set with geometric descriptor and quantum-chemical descriptors.

3. Conclusions

The above results indicate that GEP is a very promising tool for function estimation. The GEP exhibits better overall performance due to some advantages over the other techniques of converging to the global optimum and not to a local optimum. The predictive results are satisfied for regression. Therefore, it is a good approach for predicting the expected activity of drugs and aiding drug design. At the same time, the models proposed could identify and provide some insight into what structural features are related to the biological activity of these compounds and provide some instruction for further designing of the new highly active 1,4-dihydropyridine calcium channel antagonists.

4. Experimental

4.1. Data set

The structures and IC_{50} values for 45 antagonists were taken from the literature[5] and are listed in Table 1. For analysis purposes, log(1/IC_{50}) values were used as the dependent variables and are given in Table 1. The training set was used to build the model, and the test set used to evaluate its prediction ability. Leave-one-out (LOO) cross-validation was performed for the whole training set.

4.2. Calculation of the descriptors

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.).[25]

4.3. The heuristic method[25]

Once molecular descriptors are generated, the heuristic method in CODESSA was used to accomplish the preselection of the descriptors and build the linear model. Its advantages are the high speed and no software restriction on the size of the data set. 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, means error, and the F-value). The selected descriptors are shown in Table 2. The calculation formulae of Moments of inertia C, Molecular surface area, and HOMO were listed as the following.

Moments of inertia C are in the rigid rotator approximation, the principal moments of inertia of a molecule, and can be calculated as follows:

$$I_C = \sum_i m_i r_{iz}^2, \tag{3}$$

where m_i is the atomic mass and r_{iz} denotes the distance of the ith atomic nucleus from the main rotational axes z of the molecule.

Molecular surface area can be calculated as:

$$S_M = \sum_i l_i \, \mathrm{d}x. \tag{4}$$

HOMO energy can be calculated as follows:

$${}^{S}\mathrm{HOMO} = \langle {}^{\phi}\mathrm{HOMO}|\hat{F}|{}^{\phi}\mathrm{HOMO}\rangle. \tag{5}$$

${}^{\phi}$HOMO—highest occupied molecular orbital and $\hat{F}$—Fock operator.

Since the influencing $\log(1/\mathrm{IC}_{50})$ of compounds were complex and not all of them were in linear correlation with the $\log(1/\mathrm{IC}_{50})$, in order to build a more accurately predictive model, it is necessary to build a nonlinear model.

4.4. Theory of gene expression programming

Gene expression programming was invented by Ferreira in 1999[27] and was developed from genetic algorithms and genetic programming (GP). GEP uses the same kind of diagram representation of GP, but the entities evolved by GEP (expression trees) are the expression of a genome.

GEP is more simple than cellular gene progression. It mainly includes two sides: the chromosomes and the expression trees (ETs). The process of information of gene code and translation is very simple, such as a one-to-one relationship between the symbols of the chromosome and the functions or terminals they represent. The rules of GEP determine the spatial organization of the functions and terminals in the ETs and the type of interaction between sub-ETs. Therefore, the language of the genes and the ETs represents the language of GEP.[26]

(1) The GEP chromosomes, expression trees (ETs), and the mapping mechanism.

Each chromosome in GEP is a character string of fixed-length, which can be composed of gene from the function set or the terminal set. Using the elements {+, −, ∗, /, Q} as the function set and $\{a, b, c, d\}$ as the terminal set, the following is an example of GEP chromosome of length eight:

$$\begin{array}{c} 01234567 \\ Q^{*} + -abcd, \end{array} \tag{6}$$

where Q denotes the square-root function; and a, b, c, d are variable (or attribute) names. The above is referred to as Karva notation or K-expression.[26] A K-expression can be mapped into an ET following a width-first procedure. A branch of the ET stops growing when the last node in this branch is a terminal. For example, the ET shown in Figure 1 corresponds to the sample chromosome (6) and can be interpreted in a mathematical form as (7). The conversion of an ET into a K-expression is also very straightforward and can be accomplished by recording the nodes from left to right in each layer of the ET in a top-down fashion to form the string.

$$y = \sqrt{(a-b) * (c+d)}. \tag{7}$$

(2) The description of the GEP algorithm.

The purpose of symbolic regression or function finding is to find an expression that can give good explanation for the dependent variable. The first step is to choose the fitness function. Mathematically, the fitness f_i of an individual program i is expressed by the equation:

$$f_i = \sum_{j=1}^{n} \left(R - \left| \frac{P_{(ij)} - T_j}{T_j} \cdot 100 \right| \right), \tag{8}$$

where R is the selection range, $P_{(ij)}$ is the value predicted by the individual program i for fitness case j (out of n fitness cases), and T_j is the target value for fitness case j. Note that the absolute value term corresponds to the relative error. This term is what is called the precision and if the error is smaller than or is equal to the precision then the error becomes zero. Thus, for a good match the absolute value term is zero and $f_i = f_{\max} = nR$.

For some function finding problems it is important to evolve a model that performs well for all fitness cases within a certain relative error (the precision) of the correct value.

The fitness $f_{(ij)}$ of an individual program i for fitness case j is evaluated by the formula:

$$\text{If } E_{(ij)} \leqslant p, \quad \text{then } f_{(ij)} = 1; \quad \text{else } f_{(ij)} = 0, \tag{9}$$

where p is the precision and $E_{(ij)}$ is the relative error of an individual program i for fitness case j evaluated by the equation:

$$E_{(ij)} = \left| \frac{P_{(ij)} - T_j}{T_j} \cdot 100 \right|, \tag{10}$$

where $P_{(ij)}$ the value is predicted by the individual program i for fitness case j and T_j is the target value for fitness case j.

The other two evaluated functions are introduced below:

The first, mean squared error E_i of an individual program i is evaluated by the equation:

$$E_i = \frac{1}{n} \sum_{j=1}^{n} (P_{(ij)} - T_j)^2, \tag{11}$$

where $P_{(ij)}$ is the value predicted by the individual program i for fitness case j (out of n fitness cases) and T_j is the target value for fitness case j.

For a perfect fit, $P_{(ij)} = T_j$ and $E_i = 0$. So, the E_i index ranges from 0 to infinity, with 0 corresponding to the ideal. The advantage of this kind of fitness function is that the system can find the optimal solution for itself.[27]

The second major step consists in choosing the set of terminals T and the set of functions F to create the chromosomes. In this problem, the terminal set consists obviously of the independent variable, that is, $T = \{a\}$. The choice of the appropriate function set is not so obvious, but a good guess can always be done in order to include all the necessary functions.

The third major step is to choose the chromosomal architecture, that is, the length of the head and the number of genes.

The fourth major step in preparing to use gene expression programming is to choose the linking function.

And finally, the fifth major step is to choose the set of genetic operators that cause variation and their rates. This process is repeated for a pre-specified number of generations or until a solution has been found. In GEP, individuals are often selected and copied into the next generation based on their fitness, as determined by roulette-wheel sampling with elitism,[28] which guarantees the survival and cloning of the best individual to the next generation. Variation in the population is introduced by applying one or more genetic operators to selected chromosomes, including: crossover, in which two parent chromosomes are randomly chosen and paired to exchange some elements between them. There are two kinds of crossover: one-point and two-point crossover, working in the same fashion as in the canonical GAs;[28] mutation, in which the symbols at any position in a chromosome are subjected to a random change according to a certain probability; rotation, in which two subparts of the element sequence in a chromosome are rotated with respect to a randomly chosen point. Note that all of these operations upon the coding sequence of a chromosome usually drastically reshape the corresponding ETs.

The flowchart of a gene expression algorithm (GEA) is shown in Figure 2. The process begins with the random generation of the chromosomes of the initial population. Then the chromosomes are expressed and the fitness of each individual is evaluated. The individuals are then selected according to fitness to reproduce with modification, leaving progeny with new traits. The individuals of this new generation are, in their turn, subjected to the same developmental process: expression of the genomes, confrontation of the selection environment, and reproduction with modification. The process is repeated for a certain number of generations or until a solution has been found.

To evaluate the ability of GEP, correlation coefficient was proposed:

$$C_i = \frac{\mathrm{Cov}(T, P)}{\sigma_t \cdot \sigma_p}, \tag{12}$$

where $\mathrm{Cov}(T, P)$ is the covariance of the target and model outputs; and σ_t and σ_p are the corresponding standard deviations.

All calculation programs implementing GEP were written in GepModel.[18] The GEP software package was programmed by C^{++} language and running operating system on a Pentium IV with 256 M RAM.

Acknowledgments

The authors thank the Gepsoft Team for providing the free Gepsoft software. We also thank R Development Core Team for affording the free R1.7.1 software. The authors thank the National Natural Science Foundation of China (NSFC) Fund (No. 20305008) for supporting this project too.

References and notes

1. Ana, B. B.; Rosa, M. A.; Rosa, M. J.; Wolfgang, W. *Forensic Sci. Int.* **2006**, *156*, 23.
2. Godfraind, T.; Salomone, S. *J. Cardiovasc. Pharmacol.* **1997**, *30*, S1.
3. Van Zwieten, P. A. *Blood Press* **1998**, *7*, 5.
4. Nayler, W. G. *J. Clin. Basic Cardiol.* **1999**, *2*, 155.
5. Costa, M. C. A.; Gaudio, A. C.; Takahata, Y. *THEOCHEM* **1997**, *394*, 291.
6. Zamponi, G. W.; Stotz, S. C.; Staples, R. J.; Andro, T. M.; Nelson, J. K.; Hulubei, V.; Blumenfeld, A.; Natale, N. R. *J. Med. Chem.* **2003**, *46*, 87.
7. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328.

8. Takahata, Y.; Costa, M. C. A.; Gaudio, A. C. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 540.
9. Viswanadhan, V. N.; Mueller, G. A.; Basak, S. C.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 505.
10. Hemmateenejad, B.; Miri, R.; Akhond, M.; Shamsipur, M. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 91.
11. Schleifer, K. J.; Tot, E. *Quant. Struct.-Act. Relat.* **2002**, *21*, 239.
12. Yao, X. J.; Liu, H. X.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Panaye, A. J.; Doucet, P.; Fan, B. T. *Mol. Pharmacol.* **2005**, *2*, 348.
13. Alice, B.; Daniele, P.; Patrick, D.; Anna, M. B. *Bioorg. Med. Chem.* **2005**, *13*, 5330.
14. Andrei, I. K.; Igor, A. S.; Mark, T. Q. *Bioorg. Med. Chem.* **2006**, *14*, 352.
15. Li, W.; Tang, Y.; Zheng, Y. L.; Qiu, Z. B. *Bioorg. Med. Chem.* **2006**, *14*, 601.
16. Ozlem, T.; Keskin, M. E. *J. Appl. Sci.* **2005**, *5*, 508.
17. Baykasoglu, A.; Dereli, T.; Tanis, S. *Cem. Concr. Res.* **2004**, *34*, 2083.
18. http://www.gepsoft.com/gepsoft.
19. Handbook of Chemistry and Physics, CRC Press, Cleveland OH, 1974, p F-112.
20. Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chim. Acta* **1987**, *199*, 99.
21. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; J. Wiley & Sons: New York, 2000.
22. Csizmadia, I. G. *Theory and Practice of MO Calculations on Organic Molecules*; Elsevier: Amsterdam, 1976.
23. Clare, B. W. *Theor. Chim. Acta* **1994**, *87*, 415.
24. Kiyoshi, H.; Yoshikatsu, M.; Kimito, F. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306.
25. Katritzky, A. R; Lobanov, V. S; Karelson, M.; Reference Manual, Version 2.0, 1994.
26. Ferreira, C. *Gene Expression Programming in Problem Solving. Soft Computing and Industry-Recent Applications*; Springer-Verlag: Berlin, 2002, pp 635–654.
27. Ferreira, C. *Complex Systems* **2001**, *13*, 210.
28. Mitchell, M. *An Introduction to Genetic Algorithms Complex Adaptive Systems*; MIT Press: Cambridge, MA, 1996.