Journal of Computational Science & Engineering 1 (2011) 22-31

Journal of Computational Science & Engineering
Available at http://icse.qdu.edu.cn ISSN 1710-4068

Study the chromatographic hydrophobicity index based on gene expression programming

Hongzong Si[1*], Hua Gao[1], Xiaojun Yao[2] and Zhide Hu[2]

[1]*Institute for Computational Science and Engineering Qingdao University, Qingdao, Shandong 266071, China;*
[2]*Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, China*

Article Information

Article history:
Received 25 December 2010
Revised 6 January 2011
Accepted 15 January 2011
Available online 25 January 2011

Keywords:
Chromatographic hydropho-bicity index
Gene expression programing
Quantitative structure-property relationship
HM

Abstract

A novel nonlinear method of gene expression programming (GEP) was used to develop nonlinear models for predicting the chromatographic hydrophobicity index. The six descriptors were selected by heuristic method (HM) and were used as the inputs of the linear and nonlinear models respectively. Number of genes, the functions selecting, and the difference of training and test sets are the most important factors for setting up quantitative structure-property relationship (QSPR) model in the GEP method. Results indicate that the GEP is a reliable nonlinear method for building QSPR model.

1. Introduction

The chromatographic hydrophobicity index (CHI) is a parameter that is designed to measure lipophilicity of compounds. It is derived from the solvent strength to elute the compound from a reversed-phase high- performance liquid chromatography (HPLC) column. The absolute magnitudes of the parameter CHI dependents on the values assigned to the set of standards of calibrate the gradient [1]. Therefore, CHI was used for the retention time in a calibrated generic gradient of HPLC experiment [2] and was introduced as a high-throughput method for lipophilicity determination. The relationship of CHI with retention time is $CHI = At_R + B$. Here, A and B are the constants of a linear plot of CHI values which is a determined set of standards against their gradient retention times. The constants, A and B, are used to calibrate the gradient system. t_R is the retention time.

CHI is deduced from the retention time. It not only reflects the lipophilicity of the compound but also approximates the concentration of organic phase of an equal distribution of between the mobile phase and stationary phase [3].

The calculated descriptors of 32 compounds based on the software CODESSA were used to predict the CHI by the HM and GEP. The GEP is a novel nonlinear regression method. It has a stronger generalization performance [4, 5]. In our previous works, the GEP was used to construct

* Corresponding author. E-mail: *sihz03@126.com*

and obtained good QSAR models [6, 7]. In order to test the stability of QSPR model and the repetition, we investigated the key indices: the number of genes, the selection functions, the difference of training and test set. The QSPR has the following advantages: (1) this method only need the structure of compounds; (2) once a model was built, it is very easy to get the CHI values. In this paper, the GEP method was used in this investigation to build the QSPR model.

Table 1 CHI values of validation mixture by LC/MS and LC/UV at the three pH values (differences (△) between the two kinds of measurements), the physicochemical meanings of descriptors

no.	compound name	pH 2.0			pH 7.4			pH 10.5		
		CHI^{MS}	CHI^{UV}	△	CHI^{MS}	CHI^{UV}	△	CHI^{MS}	CHI^{UV}	△
1	acetazolamide	23.22*	23.02*	0.20	11.54*	11.54*	0.00	-48.05	-48.05	0.00
2	allopurinol	7.49	7.23*	0.26	-14.55	-14.55	0.00	-43.62	-43.62	0.00
3	bendroflumethiazide	72.9	72.88*	0.01	75.86	75.86	0.00	67.42	67.38	0.04
4	benzocaine	58.82*	58.76*	0.07	63.26*	63.26*	0.00	64.09	64.09	0.00
5	benzthiazide	67.52	67.48*	0.03	56.52	56.52	0.00	44.32	44.32	0.00
6	betamethasone	59.65	59.59*	0.06	59.21	59.21	0.00	60.46	60.46	0.00
7	butamben	82.42*	82.44*	-0.02	80.80	80.80	0.00	81.44	81.44	0.00
8	butylparaben	82.42	82.44*	-0.02	79.9*	79.9*	0.00	47.55	47.55	0.00
9	carbamazapine	60.48	60.42	0.06	61.46	61.46	0.00	64.09	64.09	0.00
10	carisoprodol	67.52	67.48	0.03	28.63	28.63	0.00	67.32	67.32	0.00
11	chloroquine	14.53*	14.29	0.23	84.40*	84.40*	0.00	58.44	58.44	0.00
12	chloroxylenol	82.42	82.44	-0.02	79.9	79.9	0.00	77	77	0.00
13	chlorpropamide	69.17	69.14	0.03	32.68	32.68	0.00	35.45	35.45	0.00
14	clofazimine	61.31	61.25	0.06	124.88	124.88	0.00	158.89	158.89	0.00
15	droperidol	36.47*	36.32	0.15	76.75*	76.75*	0.00	77.81	77.81	0.00
16	gemfibrozil	101.05	101.14	-0.09	65.51	65.51	0.00	51.18	51.18	0.00
17	hydrocortisone	52.61	52.52	0.09	52.02	52.02	0.00	53.6	53.6	0.00
18	hydroflumethiazide	42.68	42.55	0.13	44.82	44.82	0.00	93.14	93.12	0.00
19	iodipamide	83.66	83.69	-0.03	26.38	26.38	0.00	28.59	28.59	0.00
20	nitofurazone	34.81*	34.65	0.16	37.18*	37.18*	0.00	40.69	40.69	0.00
21	oxyphenbutazone	65.45	65.4	0.04	34.48	34.48	0.00	27.79	27.78	0.00
22	phenacemide	44.33	44.21	0.12	43.92	43.92	0.00	48.36	48.36	0.00
23	phenylbutazone	94.84	94.91	-0.07	46.17	46.17	0.00	43.11	43.11	0.00
24	prednisone	53.44*	53.35	0.09	53.82*	53.82*	0.00	55.22*	55.22*	0.00
25	primidone	37.3	37.15	0.15	36.28	36.28	0.00	39.89*	39.89*	0.00
26	tetracaine	39.37	39.23	0.14	84.4	84.4	0.00	84.66*	84.66*	0.00
27	tetracycline	55.1	55.02	0.08	48.87	48.87	0.00	51.18*	51.18*	0.00
28	trimethoprim	20.74*	20.53	0.21	42.12*	42.12*	0.00	45.13*	45.13*	0.00
29	hydroquinine	17.84	17.62	0.22	67.31	67.31	0.00	72.56*	72.56*	0.00
30	morin	50.54	50.44	0.10	30.43	30.43	0.00	-45.63*	-45.63*	0.00
31	phenyl	101.05	101.14	-0.09	99.08	99.15	0.07	99.51*	99.59*	0.08
32	progesterone	100.22	100.31	-0.09	96.09	96.09	0.00	93.14*	93.14*	0.00

* test set

2. Experimental section

2.1 Data preparation

The CHIs values of reversed-phase liquid chromatography method were collected [8] and were given in Table 1. The data set is randomly separated into a training set and a test set. The training set is used to build the model and the test set is employed to evaluate the prediction ability of the model.

2.2 Calculation of the descriptors

The numerical representation (often called molecular descriptor) of the chemical structure is the most important factors affecting the quality of the QSPR model. In the present investigation, the process of the molecular descriptors is described as below: all molecules were drawn into Hyperchem [9] and pre-optimized using MM+ molecular mechanics force field. A more precise optimization was done with semi-empirical AM1 method in MOPAC [10]. 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 [11,12] 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.); quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.). The software CODESSA, developed by Katritzky group, enables the calculation of a large number of quantitative descriptors based solely on the molecular structural information and codes chemical information into mathematical form [11, 12]. CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish molecular QSAR/QSPR models. CODESSA has been applied successfully in a variety of QSPR analyses [13-15].

2.3 The selection of the descriptors and the development of linear model by HM [11, 12]

Once molecular descriptors are generated, the HM in CODESSA was used to accomplish the pre-selection of the descriptors and build the linear model. The advantages of this model are the high speed and no software restrictions on the size of the data set. The heuristic method can either quickly gives a good estimation about what quality of correlation to expect from the data, or derive several best regression models [13-16].

2.4 The GEP theory

The GEP was invented by Ferreira in 1999 [17], and it is the natural development from genetic algorithms and the genetic programming (GP). The GEP uses the same kind of diagram representation of the GP, and the difference is that the GEP is expressed with genes.

The GEP includes chromosomes and the expression trees (ETs). The processes of gene encoding and translation are very simple. It is one-to-one relationship between the symbols of the chromosome and the functions or terminals they represent. The GEP rules determine the spatial organization of functions and terminals in the ETs and the type of the interaction between sub-ETs. Therefore, the genes and the ETs represent the GEP language [18]. Each chromosome in the GEP is a character string of the fixed-length, which can be composed of the gene from the function set or the terminal set. The elements {+, -, *, /, Q} are used as the function set and {a, b, c, d} as the terminal set. The following equation is an example of the GEP chromosome with the length eight

$$\begin{Bmatrix} 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7 \\ Q\ *\ +\ -\ a\ b\ c\ d \end{Bmatrix}, \qquad (1)$$

where Q denotes the square-root function, and a, b, c, d are variable (or attribute) names. The Eq. (1) is referred to as the Karva notation or the K-expression [18]. 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. Eq. (1) and can be expressed in a mathematical form of $\sqrt{(a-b)*(c+d)}$. ET can be directly converted into a K-expression by recording the nodes from left to right in each layer of the ET in a top-down fashion to form the string.

2.5 Development of nonlinear model by the GEP algorithm

The purpose of the symbolic regression or function is to find an expression that can give a good explanation for the dependent variable. The process of the GEP has five steps. The first step is to choose the fitting function. Mathematically, the fitness of an individual program is expressed by the equation

$$f_i = \sum_{j=1}^{n}\left(R - \left| \frac{P_{(ij)} - T_j}{T_j} \bullet 100 \right| \right), \quad (2)$$

where R is the selection range, $P_{(ij)}$ is the value predicted by the individual program i for fitting case j (out of *n* fitting cases), and T_j is the target value for the fitting case j. For some function finding problems, it is important to evolve a model that performs well for all fitting cases within a certain relative error. The fitness $f_{(ij)}$ of an individual program *i* for the fitting case *j* is formulated as

$$f(i,j) = \begin{cases} 1, E_{(i,j)} \le p \\ 0, \end{cases} \quad (3)$$

where p is the precision and $E_{(ij)}$ is the relative error of an individual program i for the fitting case j. The $E_{(ij)}$ [18] is given by

$$E_{(ij)} = \left| \frac{P_{(ij)} - T_j}{T_j} \bullet 100 \right|. \quad (4)$$

The second step consists of 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, i.e., $T = \{a\}$. The third step is to choose the chromosomal architecture, i.e., the length of the head and the number of genes. The fourth major step is to choose the linking function. The last major step is to choose the set of genetic operators that cause variation and their rates. These processes are repeated for a pre-specified number of generations until a solution is obtained. In the GEP, the individuals are often selected and copied into the next generation based on their fitness, as determined by roulette-wheel sampling with elitism [19], which guarantees the survival and cloning of the best individual to the next generation. The variation in the population is introduced by applying one or more genetic operators to select chromosomes, including crossover, mutation, and rotation.

The process begins with the random generation of the chromosomes of the initial population. The chromosomes are expressed and the fitness of each individual is evaluated. The individuals are selected according to the 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.

To evaluate the ability of the GEP, the correlation coefficient (*R*) was introduced as

$$C_i = \frac{Cov(T,P)}{\sigma_t \bullet \sigma_p} \quad (5)$$

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

2.6 The GEP implementation and the computational environment

The computing programs implementing the GEP are written in GepModel. The GEP software package is programmed in C^{++} language and performed on a Pentium IV computer with a 512 M RAM system.

Table 2 The meanings of descriptors

Descriptors	physicochemical meanings
CAN	Number of N atoms
CBR	Number of benzene rings
PPSA-3	PPSA-3 Atomic charge weighted PPSA [Zefirov's PC]
ANRC	Avg nucleoph. react. index for a C atom
RPCSSA	RPCS Relative positive charged SA (SAMPOS*RPCG) [Quantum-Chemical PC]
HDCAH	HDCA H-donors charged surface area [Quantum-Chemical PC]
HDCA-1/TMSA	HA dependent HDCA-1/TMSA [Zefirov's PC]
THCMD	Tot hybridization comp. of the molecular dipole
ASI	Average Structural Information content (order 0)
PPSA-1	PPSA-1 Partial positive surface area [Zefirov's PC]
GP	(1/6)X GAMMA polarizability (DIP)
CAB	Number of aromatic bonds
HAD	HA dependent HDSA-2/TMSA [Zefirov's PC]
HOMO-LUMO	HOMO - LUMO energy gap
RMW	Relative molecular weight
HOMOE	HOMO energy
MTICH	Max total interaction for a C-H bond
HOMO-1	HOMO-1 energy

3. Results and discussion

3.1 Results of HM

Total 660 descriptors were calculated by the CODESSA program for all the compounds. After the heuristic reduction, the pool of descriptors was reduced to 260. In every data group, 6 descriptors were selected, which were most relevant to the CHI of compounds. Meanwhile, to avoid the "over-parame- terization" of the model, an increasing of the R^2 values of less than 0.02 was chosen as the breakpoint criterion [20]. Six descriptors of 6 groups were eventually selected respectively. The meanings of descriptors were summarized in Table 2. The descriptors are the same by the HM and the GEP in every group. We obtained linear models by HM as follows (Figure 1-6).

The same descriptors were selected in two groups LC/MS and LC/UV by HM. The experimental data of two groups are very similarity (Table 1). The calculated values of CHI are in good agreement with experimental data respectively. The two groups LC/MS and LC/UV data of calculation are much closed to each other.

The equations of (8) and (9) showed the same results with the pH=7.4 and 10.5. For the nearly same original data, the same models of MS and UV methods of pH=7.4 and 10.5 were found respectively.

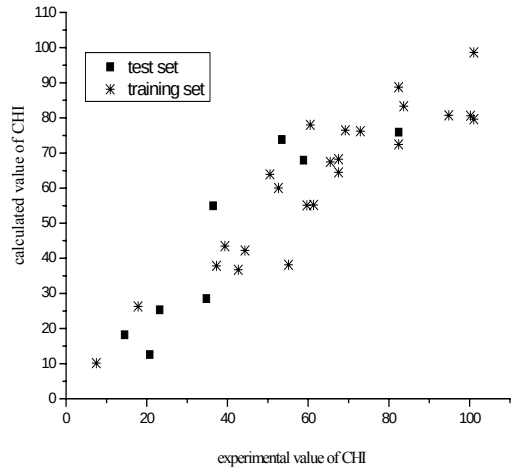
The good QSPR models were found for the exam data of LC/MS and LC/UV methods. Table 1 shows the little difference of CHI between the LC/MS and LC/UV methods. Basis on the experimental results, the predicted values are in good agreement with it. However, QSPR models by HM for the exam data of LC/MS and LC/UV methods get nearly common results.

$$CHI^{MS}_{(pH=2.0)} = 1.74\times10^{2} - 1.84\times10^{1}CAN + 1.65\times10^{1}CBR - 7.53\times10^{0}(PPSA\text{-}3) - 9.66\times10^{3}ANRC + 9.27\times10^{0}RPCSSA + 1.39\times10^{0}HDCAH \quad (6)$$

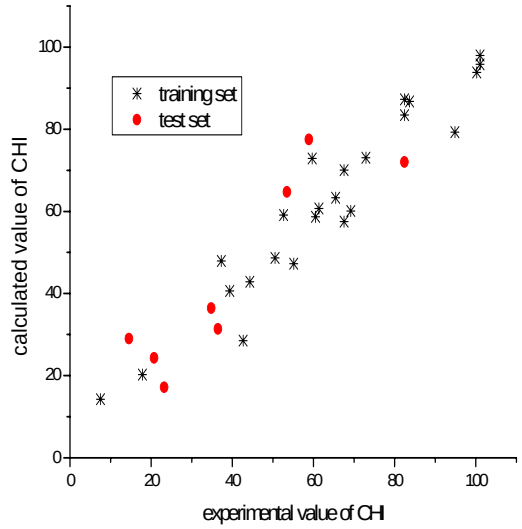
$$CHI^{UV}_{(pH=2.0)} = 1.74\times10^{2} - 1.84\times10^{1}CAN + 1.65\times10^{1}CBR - 7.56\times10^{0}(PPSA\text{-}3) - 9.7\times10^{3}ANRC + 9.31\times10^{0}RPCSSA + 1.4\times10^{0}HDCAH \quad (7)$$

$$CHI^{MS,\,UV}_{(pH=7.4)} = 1.0\times10^{3} - 1.36\times10^{3}HAD - 3.62\times10^{1}(HOMO - LUMO) - 5.62\times10^{0}RMW - 4.76\times10^{1}HOMOE - 6.31\times10^{1}MTICH + 1.83\times10^{1}(HOMO\text{-}1) \quad (8)$$

$$CHI^{MS,\,UV}_{(pH=10.5)} = 2.47\times10^{2} - 5.74\times10^{3}(HDCA\text{-}1/TMSA) + 3.47\times10^{1}THCMD - 2.96\times10^{2}ASI - 2.05\times10^{-1}(PPSA\text{-}1) + 1.35\times10^{-3}GP - 1.96\times10^{0}CAB \quad (9)$$

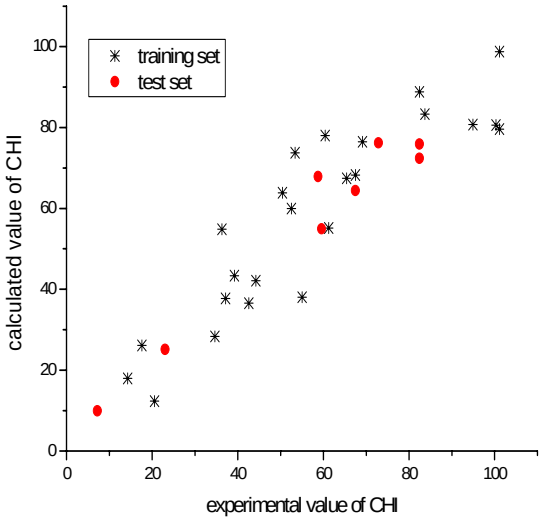


1a

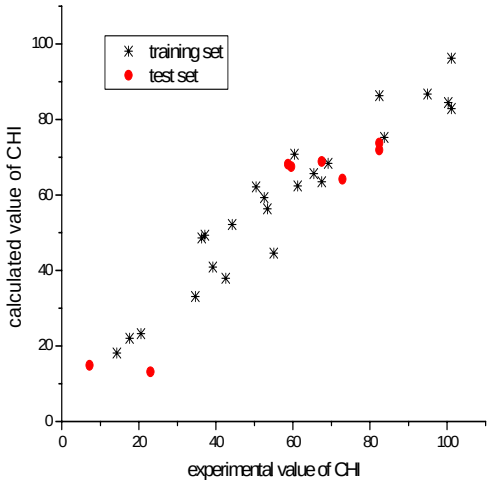


1b

Figure 1. Predicted CHI values vs. experi- mental values (pH=2.0, MS method, 1a is the model of HM; 1b is the model of GEP).

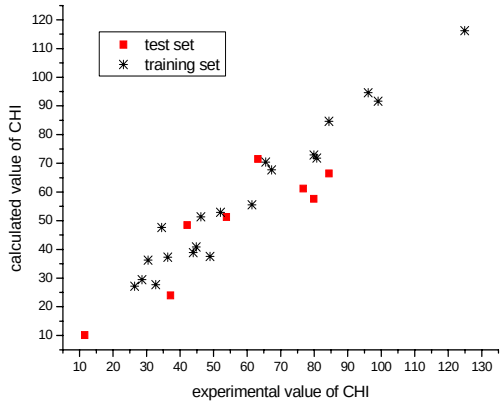


2a

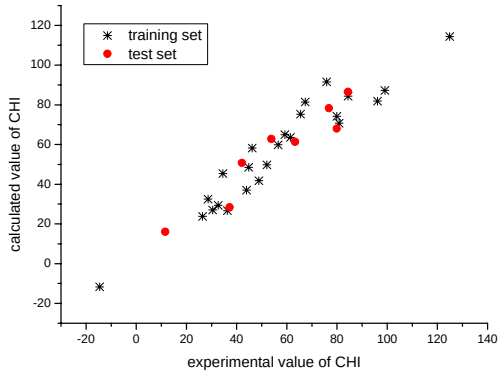


2b

Figure 2. Predicted CHI values vs. experi- mental values (pH=2.0, UV method, 2a is the model of HM; 2b is the model of GEP.).

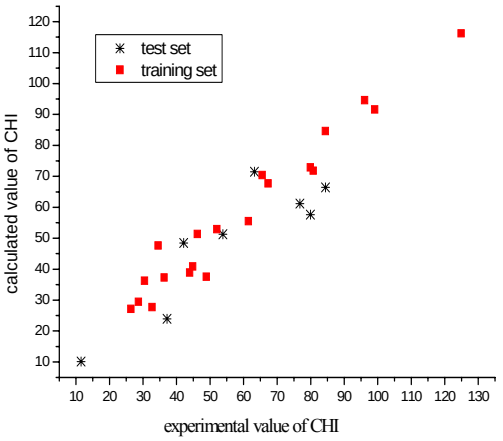


3a

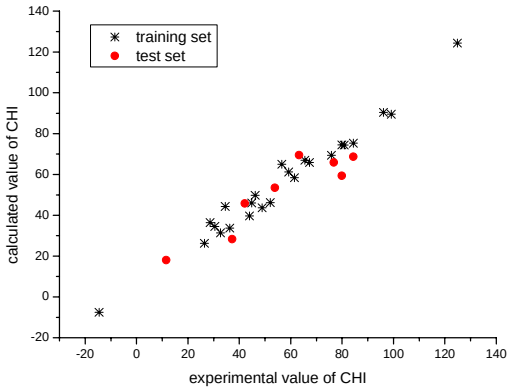


3b

Figure 3. Predicted CHI values vs. experi- mental values (pH=7.4, MS method, 3a is the model of HM; 3b is the model of GEP.).

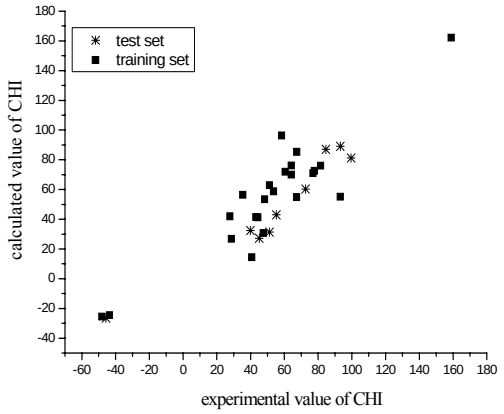


4a

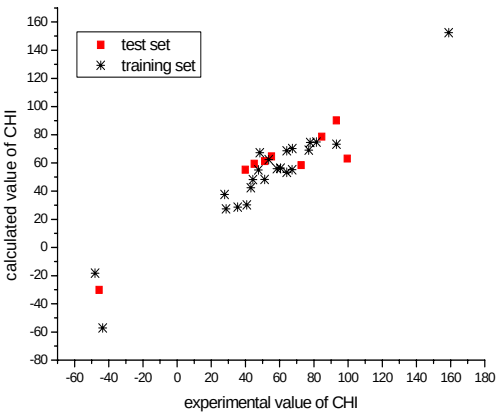


4b

Figure 4. Predicted CHI values vs. experi- mental values (pH=7.4, UV method, 4a is the model of HM; 4b is the model of GEP.).

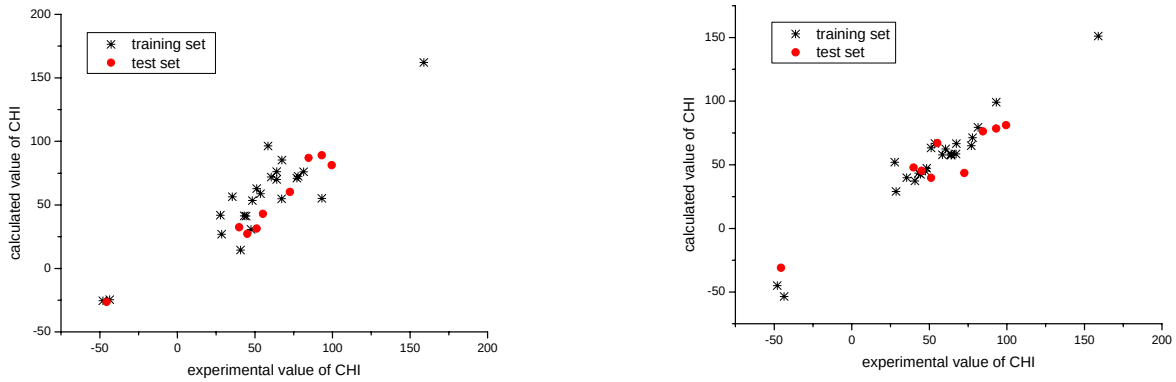


5a



5b

Figure 5. Predicted CHI values vs. experi- mental values (pH=10.5, MS method, 5a is the model of HM; 5b is the model of GEP.).



6a 6b

Figure 6. Predicted CHI values vs. experi- mental values (pH=10.5, UV method, 6a is the model of HM; 6b is the model of GEP.).

3.2 Results of the GEP

The software automatic problem solver (APS) [19] 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 [21]. All descriptors of three types of pH values were all selected. Four QSPR models were constructed by HM. In order to obtain the best QSPR model, the GEP was used to build QSPR models with the same descriptors respectively (Figure 1-6).

The selection of training set and test set will affect the QSPR model (Figures 1, 2). Equations (10) and (11) were obtained by the GEP method. The same functions (+, -, *, /, cos, fabs) were used in equations (10) and (11). At the same time, the number of genes is 5 for two equations. We only changed the different compounds entered into training or test set. In equation (10), the test set contained 8 compounds 1, 4, 7, 11, 15, 20, 24 and 28. In equation (11), the test set contained 8 compounds 1, 2, 3, 4, 5, 6, 7 and 8. Different training set and test set will produce the different correlation coefficients. However, higher correlation coefficient of training set often leads to the lower correlation coefficient of test set. Changing the compounds in test sets will also lead to the mean square error (MSE) increase rapidly. Equation (11) has higher correlation coefficient of test set, i.e. the model will have stronger generalization performance. The MSE are reduced greatly of equation (11) than that of equation (10).

More number of genes will increase the generalization performance of QSPR model (Figure3, 4). In equations (12) and (13), the same test set contained 8 compounds 1, 4, 8, 11, 15, 20, 24 and 28, and the same functions (+, -, *, /, pow, fabs) were selected. The number of genes was 8 in equation (12), which was increased to 17 in equation (13). From equation (12) and (13), the coefficient of test set increase greatly with the number of gene increasing. It shows that the model of (13) has powerful prediction ability. At one time, the MSE decreased too.

Changing the function affect the building of different QSPR models (Figure5, 6). In the GEP, the same condition will get the same QSPR model. In equations (14) and (15), the same test set contained 9 compounds 24, 25, 26, 27, 28, 29, 30, 31 and 32. 5 genes were selected in two equations. In equations (14), functions (+, -, *, /, cos) were selected, at the same time, in equations (15), functions (+, -, *, /, cos, fabs) were selected. Increasing one function can improve the correlation coefficient of training set, while the correlation coefficient of test set decreasing and the MSE of test set increasing.

$$\begin{aligned}&CHI_{(pH=2.0)}^{MS}=(6.7-\cos x_1)\,|\,x_2-x_3\,|-11x_1+\\&|\,x_4-\frac{72}{x_3}-9.6+\cos x_5\,|+\frac{x_5}{x_3}+x_3x_4+\\&\cos(x_6x_4\,7.8x_4+9.8x_6+76.44)+\\&|\,(50.96+x_1-x_2)(x_1x_5-x_1)\,|\end{aligned}\tag{10}$$

$$\begin{aligned}&CHI_{(pH=2.0)}^{UV}=\left|-9.98\cos\left(\left(x_3-x_5\right)*x_6\right)-9.98\right|\\&-1.47x_1*\left|-8.84-\cos x_5\right|+\\&\cos\left(\left(x_2+x_4\right)*x_3\right)*\left(x_6-x_3+x_2x_3\right)+x_2-\\&\cos x_3+x_5-x_3+\left\|9.92(-7.21-x_2)\right|+x_2\right|\end{aligned}\tag{11}$$

In equations 10 and 11, x_1 to x_6 repre- sent CAN, CBR, PPSA-3, ANRC, RPCSSA and HDCAH respectively.

$$\begin{aligned}&CHI_{(pH=7.4)}^{MS}=x_2\left(\frac{1}{x_1-1.26x_5x_1}-1\right)+x_3+\\&|\,x_4x_5-x_3-18.37\,|+\\&\left|\left(\frac{6.98}{x_3}\right)^{(2x_5-1.99)}\right|+\left|x_3^{\left|\frac{x_6}{x_4^2}+x_3\right|}\right|+\\&13.42-x_5*\left(\left(x_1x_5\right)^{x_4}+1\right)+\\&(x_1+x_2)*\left(\frac{x_1x_2x_5}{x_3}\right)-(x_1x_3x_6)*\\&(x_2+x_6-x_3)+\left\|\frac{x_3}{x_1x_2}\right|+x_6\right|\end{aligned}\tag{12}$$

$$\begin{aligned}&CHI_{(pH=7.4)}^{UV}=2x_6-x_2\left(x_5+1\right)+\\&x_3\left(x_1x_6+\left|x_6\right|^{-1.88}-1\right)+\\&4.29^{(9.32+x_2^{-1.85}-x_2)}+\left|(x_5+x_3)*x_1x_3-x_4-8.27\right|+\\&x_1\left(3+\frac{x_4}{x_3}+\frac{1}{x_3+x_5}\right)+\\&\frac{x_6+2x_5}{x_4}-x_4\left(\frac{1}{x_5}-x_5-3\right)+x_5^{x_6}+\\&\frac{3.3x_4}{x_1x_2(x_1+x_6)}+\left(\frac{x_6}{5.96}\right)^{\left|\frac{3.4}{x_3+x_4}\right|}+\\&\left|2x_1x_3x_5+x_2-9.85\right|+\left|8.24-x_2\right|+6.34+\left|x_4\right|+\left|x_1\right|\end{aligned}\tag{13}$$

In equations 12 and 13, x_1 to x_6 repre- sent HAD, HOMO-LUMO, RMW, HOMOE, MTICH and HOMO-1 respectively.

$$\begin{aligned}&CHI_{(pH=10.5)}^{MS}=8.79-\cos\left(\frac{x_5-0.44}{x_1x_3}\right)+\\&\frac{x_5*\cos x_1*(x_2-x_1-x_3)}{x_3}\\&+\frac{9.85}{x_3+9.95x_1*(x_2+x_6)}+\frac{x_2*\cos\left(\frac{-3.42}{x_1}\right)}{x_3*\cos x_4}-\\&\frac{3.27*(x_3-x_1)}{x_3+x_6-6.71}\end{aligned}\tag{14}$$

$$\begin{aligned}&CHI_{(pH=10.5)}^{UV}=-3\cos\left(x_5(x_4+1)-x_1+6.8-x_3\right)+\\&\frac{7.8}{x_3}+\left|x_2-x_3\right|*x_2x_6\\&-\frac{7.2}{\left(\frac{x_5}{x_4}-x_4\right)*x_1x_3}+\left|\frac{x_2}{\cos\left(2.3x_5+x_1\right)}+10\right|-\frac{3.3x_3}{\cos x_4}\end{aligned}\tag{15}$$

In equations 14 and 15, x_1 to x_6 repre- sent HDCA-1/TMSA, THCMD, ASI, PPSA-1, GP and CAB respectively.

Changing the function affect the building of different QSPR models (Figure5, 6). In the GEP, the same condition will get the same QSPR model. In equations (14) and (15), the same test set contained 9 compounds 24, 25, 26, 27, 28, 29, 30, 31 and 32. 5 genes were selected in two equations. In equations (14), functions (+, -, *, /, cos) were selected, at the same time, in equations (15), functions (+, -, *, /, cos, fabs) were selected. Increasing one function can improve the correlation coefficient of training set, while the correlation coefficient of test set decreasing and the MSE of test set increasing.

In above discussions we found that the GEP has powerful prediction than the HM. The correlation coefficients in the GEP are higher than that of the HM. The MSE in the GEP are lower than that of the HM.

4. Conclusions

QSPR study was applied to a set of compounds of CHI. QSPR models were obtained using the HM and the GEP. In the GEP, different training set and test set will produce the different correlation coefficient. Increasing the number of gene will greatly improve the coefficient of test set. Increasing function can improve the correlation coefficient of training set, while the correlation coefficient of test set decreasing and the MSE of test set increasing. The GEP has powerful prediction than the HM. The ultimate test of these models is their ability to predict CHI for newly reported compounds.

Acknowledgements

The authors are grateful to the Gepsoft Team for providing the gepsoft software. We also thank the National Natural Science Found- ation of China (20703027), the Qingdao University Research Fund for financial support (06300537) for financial supports and the Educational Commission of Shandong Province (J09LB06, J09LF16, J10LB06).

References

[1] K. Valko, C. Bevan, D. Reynolds, Anal. Chem., 69 (1997) 2022-2029.
[2] K. Valkó, P. Slégel, J. Chromatogr., 631 (1993) 49-61.
[3] J. Bartalis, F.T. Halaweish, J. of Chromatogr. B. 818 (2005) 159-165.
[4] Ö. Terzi, M.E. Keskin, J. Appl. Sci. 5 (2005) 508-512.
[5] A. Baykasoglu, T. Dereli, S. Tanış, Cement and Concrete Res., 34 (2004) 2083-2090.
[6] H.Z. Si, K.J. Zhang, Z.D. Hu, B.T. Fan, QSAR Comb. Sci., 26 (2007) 41-50.
[7] H.Z. Si, T. Wang, K.J. Zhang, Y.B. Duan, S.P. Yuan, A.P. Fu, Z.D. Hu, Anal. Chim. Acta., 591 (2007) 255-264.
[8] G. Camurri, A. Zaramella, Anal. Chem., 73 (2001) 3716-3722.
[9] HyperChem 4.0, Hypercube, 1994.
[10] J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
[11] A.R. Katritzky, L V.S. obanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0, 1994.
[12] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev., 24 (1995) 279-287.
[13] H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Talanta., 71 (2007) 258-263.
[14] H.Z. Si, T. Wang, K.J. Zhang, Z.D. Hu, B.T. Fan, Med. Chem., 14 (2006) 4834-4841.
[15] R.J. Hu, H.X. Liu, R.S. Zhang, C.X. Xue, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Talanta., 68 (2005) 31-39.
[16] A.R. Katritzky, R. Petrukhin, R. Jain, MJ. Karelson, J. Chem. Inf. Comput. Sci. 41 (2001) 1521-1530.
[17] C. Ferreira, Gene Expression Programming in Problem Solving. Soft Computing and Industry-Recent Applications, Springer-Verlag, 2002, 635-654.
[18] C. Ferreira, Complex Syst., 13 (2001) 87-129.
[19] M. Mitchell, An Introduction to Genetic Algorithms Complex Adaptive Systems. MIT Press, 1996.
[20] http://www.gepsoft.com/gepsoft
[21] F. Luan, H.Z. Si, H.T. Liu, Y.Y. Wen, X.Y. Zhang, SAR QSAR Environ Res., 19 (2008) 465-479.