



## Original article

## Predicting the activity of drugs for a group of imidazopyridine anticoccidial compounds

Hongzong Si<sup>a,\*</sup>, Ning Lian<sup>b</sup>, Shuping Yuan<sup>a</sup>, Aiping Fu<sup>a</sup>, Yun-Bo Duan<sup>a</sup>, Kejun Zhang<sup>c</sup>, Xiaojun Yao<sup>d</sup><sup>a</sup> Institute for Computational Science and Engineering, Laboratory of New Fibrous Materials and Modern Textile, The Growing Base for State Key Laboratory, Qingdao University, Ningxia Road 308, Qingdao, Shandong 266071, China<sup>b</sup> School of Chemistry and Chemical Engineering, Jiangsu Teachers' University of Technology, Changzhou, Jiangsu 213001, China<sup>c</sup> Department of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China<sup>d</sup> Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, China

## ARTICLE INFO

## Article history:

Received 17 July 2008

Received in revised form

26 March 2009

Accepted 27 April 2009

Available online 5 May 2009

## Keywords:

Anticoccidial drug

Gene expression programming

cGMP-dependent protein kinase

Quantitative structure–activity relationships

## ABSTRACT

Gene expression programming (GEP) is a novel machine learning technique. The GEP is used to build nonlinear quantitative structure–activity relationship model for the prediction of the IC<sub>50</sub> for the imidazopyridine anticoccidial compounds. This model is based on descriptors which are calculated from the molecular structure. Four descriptors are selected from the descriptors' pool by heuristic method (HM) to build multivariable linear model. The GEP method produced a nonlinear quantitative model with a correlation coefficient and a mean error of 0.96 and 0.24 for the training set, 0.91 and 0.52 for the test set, respectively. It is shown that the GEP predicted results are in good agreement with experimental ones.

Crown Copyright © 2009 Published by Elsevier Masson SAS. All rights reserved.

## 1. Introduction

Coccidiosis is a parasitic disease which is the major cause of morbidity and mortality in the poultry industry. The parasites multiply in the intestinal tract and cause tissue damage. This invasion results in the interruption of feeding, digestive processes, nutrient absorption, dehydration, blood loss, and increased susceptibility to other disease agents [1]. However, from veterinarian and economical points of view coccidiosis is a protozoan disease that costs the U.S. poultry industry about \$400 million annually. The worldwide cost for preventing it need at about \$800 million. At present, the best way of controlling coccidiosis is through the use of anticoccidial drugs. In recent years, pharmaceutical companies have not brought new anticoccidials to market. However, with the rise in drug resistance shown by the coccidia, new methods are becoming increasingly important to discovery new drugs [2]. Finding a new kind of broad spectrum drugs for therapy coccidiosis are the key step in future. Literature reports a potential target of cGMP-dependent protein kinase (PKG) [3] for designing these kinds of drugs. To improve the efficiency of drug

design, high-throughput and other artificial intelligence methods are used.

Quantitative structure–activity relationships (QSARs) have become an efficient tool for reducing the time and resources required for drug discovery. QSAR techniques based molecular descriptors, which are calculated by computer and are numerical series useful chemical information. The descriptors can be correlated statistically with biological properties or even physicochemical properties. QSAR techniques have been proven successful in the discovery of antimicrobial agents for chemotherapy, including antibacterial, anti-parasitic, and other antimicrobial compounds [4–9].

The machine learning method plays a key role in constructing QSAR model. Good QSAR model can give satisfied results, which will help experiments, such as predicting the property of new compounds and drug designing. In order to get better QSAR model, gene expression programming (GEP) is proposed. The GEP is a novel nonlinear regression method [10].

In the GEP, the implementation of different genetic operators is extremely simple because of the existence of a truly functional and autonomous genome. Systems with different evolutionary behaviors can be easily simulated in the GEP. Therefore, the GEP has been successfully used to predict the evaporation estimation [11] and the cement strength [12].

\* Corresponding author. Tel.: +86 532 85950786; fax: +86 532 85950768.

E-mail address: [sihz03@126.com](mailto:sihz03@126.com) (H. Si).

In the present work, the GEP is utilized to set up the imidazopyridine anticoccidial compounds QSAR model based on the descriptors. These descriptors are calculated from the molecular structures by the software CODESSA. Then four descriptors are selected as inputs by the heuristic method (HM). In order to investigate the influence of different descriptors on the  $IC_{50}$  of imidazopyridine anticoccidial compounds, the HM is used to build several multivariable linear models. Based on this, we developed a new QSAR model to explore the  $IC_{50}$  of the compounds with diverse structures. It is shown that the GEP predicted results are better than those of HM in both training set and test set. To our knowledge, it is the first time that the GEP method is used for predicting the  $IC_{50}$  of the imidazopyridine anticoccidial compounds on the basis of the molecular structural descriptors.

## 2. Results and discussion

### 2.1. Results of HM

Totally 473 descriptors are calculated by the CODESSA program for all the compounds. After the heuristic reduction, the pool of descriptors is reduced to 218. Finally, 4 descriptors are selected in this study (Fig. 1). The predicted results for  $IC_{50}$  based on four descriptors are listed in Table 1. These set of descriptors are most relevant to the  $IC_{50}$  of compounds and show the affecting degree of different descriptors for  $IC_{50}$ . Linear and nonlinear models with three descriptors are built respectively. Meanwhile, to avoid the “over-parameterization” of the model, an increase of the  $R^2$  values of less than 0.02 is chosen as the breakpoint criterion. Four descriptors are eventually selected. The detailed description of models based on compounds in the training set is summarized in Tables 2 and 3. The QSAR model of  $IC_{50}$  is built by HM as follows (Fig. 2)

$$IC_{50} = 156.00 + 34.80MCICH + 0.84NNA + 1.69MCICN + 16.20MBCMO$$

By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the activity of the compounds and understand which interactions play an important role in the  $IC_{50}$  of these compounds.

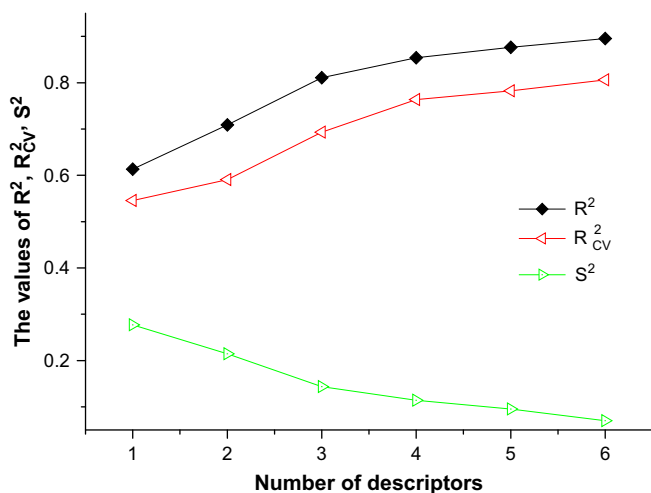


Fig. 1. Influence of the number of descriptors on the correlation coefficient ( $R^2$ ), the cross-validation correlation coefficient ( $R^2_{cv}$ ) and the square error ( $S^2$ ).

Min Coulombic interaction for a C–H bond (MCICH) and Min Coulombic interaction for a C–N bond (MCICN) are electrostatic descriptors and have positive coefficients. In this group of compounds, all molecules involve C, H and N atoms. The Coulombic interaction in the molecule is selected in the QSAR model. For electrostatics, Coulomb's law states that the direct force  $F$  of point charge  $q_1$  on point charge  $q_2$ , when the charges are separated by a distance  $r$ , is given by  $F = k_0 q_1 q_2 / r^2$ , where  $k_0$  is a constant of proportionality whose value depends on the units used for measuring  $F$ ,  $q$ , and  $r$ . The MCICH and MCICN show the minimum interaction of C–H and C–N respectively. To lower the interaction, the distance of two atoms is the key step. Due to the positive sign of the regression coefficient, the higher the descriptor value the lower its anticoccidial effect. These two descriptors indicate the importance of the intramolecular electronic effects on the intermolecular electrostatic interactions (including multipole interactions) of a molecule in determining the  $IC_{50}$  of the imidazopyridine anticoccidial compounds.

Number of N atoms (NNA) is a constitutional descriptor. For the positive coefficient, few N atoms in the compound will increase the activity. However, in this kind of drugs, the N atoms are present in all compounds. From Fig. 3 we can see that the predicted values of compounds 5, 10 and 38 are not satisfactory. Table 1 shows that the compounds 5, 10 and 38 have more N atoms than others. It is consistent with the selection of descriptor.

The quantum chemical descriptor max bonding contribution of an MO (MBCMO) is related to the strength of intramolecular bonding interactions and characterizes the stability of the molecules [13]. The coefficient of MBCMO is positive. Increasing the value of coefficient of MBCMO will lower the  $IC_{50}$ . To increase the activity of this group of compounds, the value of MBCMO should be decreased.

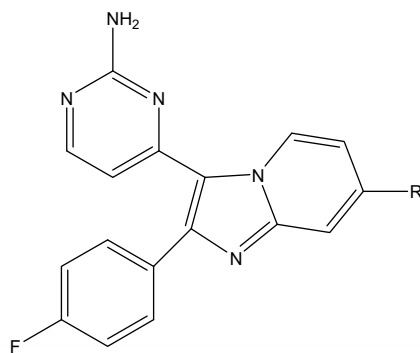
A major challenge in the development of multiple regression equations is to avoid the possible multicollinearity of molecular descriptor scales. Four descriptors are tested. Table 3 lists the correlation coefficients between the descriptors' scales, which are involved in the current four-parameter model. Table 3 demonstrates that all the descriptors are strongly orthogonal, which reflects the statistical reliability of the model.

### 2.2. Results of GEP

The software automatic problem solver (APS) [14] is 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  $R$ -square of 0.92 is obtained (Fig. 4). The C++ function is converted into the equation

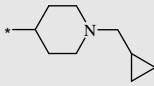
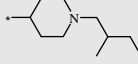
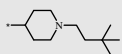
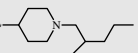
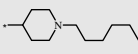
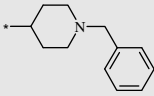
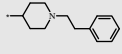
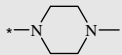
$$IC_{50} = x_1^2 + 2x_1 - x_2 - \frac{7.25x_2}{x_1^2 - 4.41x_1} - \frac{x_1x_2}{4.3x_3} + (x_1^2x_3 + x_1x_4) \lg x_4^2 + x_4x_1(1 + \lg(x_4 - 8.2)) + \frac{x_1 - x_4}{80.59x_3 - x_3^2 - 1619.75}$$

where  $x_1$ ,  $x_2$ ,  $x_3$  and  $x_4$  represent the descriptors of MCICH, NNA, MCICN and MBCMO respectively. The processes of evolving model play a key role in constructing a satisfied QSAR model. The parameters included in this model are listed in Table 4. According to the results of  $R^2$  and  $S^2$  change, the parameters are adjusted correspondingly. Functions will greatly influence the results. Finally several functions ('+', '-', '\*', '/', 'log' and 'sqrt') are selected. Head size and the number of genes are 8 and 10. Figs. 3 and 4 show that most of the predicted values are very close to the experimental ones but a few compounds have less fitting with experimental values.

**Table 1**Experimental and calculated IC<sub>50</sub> of the imidazopyridine anticoccidial compounds (HM and GEP).

Compound	R	Exp.	HM		GEP	
			Pred.	Resid.	Pred.	Resid.
1*	CH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.110	0.382	0.272	0.348	−0.238
2	CH <sub>2</sub> NHC(CH <sub>3</sub> ) <sub>3</sub>	0.350	−0.104	−0.454	−0.222	−0.572
3	CH <sub>2</sub> N(CH <sub>3</sub> )CH <sub>2</sub> CH <sub>2</sub> OH	0.500	0.454	−0.046	0.412	−0.088
4	CH <sub>2</sub> N(CH <sub>3</sub> )CH <sub>2</sub> CH <sub>2</sub> OCH <sub>3</sub>	0.690	0.430	−0.260	0.386	−0.304
5*	CH <sub>2</sub> N(CH <sub>3</sub> )CH <sub>2</sub> C≡CH	0.650	0.141	−0.509	0.043	0.607
6	CH <sub>2</sub> N(CH <sub>3</sub> )CH <sub>2</sub> C≡N	1.330	1.095	−0.235	1.578	0.248
7*	CH <sub>2</sub> N(CH <sub>2</sub> CH <sub>3</sub> ) <sub>2</sub>	0.170	−0.043	−0.213	−0.058	0.228
8		2.200	1.384	−0.816	1.932	−0.268
9	C(=O)NH <sub>2</sub>	1.800	2.353	0.553	2.106	0.306
10*	C(=NH)N(CH <sub>3</sub> ) <sub>2</sub>	2.500	2.920	0.420	3.649	−1.149
11	C(=O)NHCH <sub>3</sub>	3.100	2.688	−0.412	3.168	0.068
12	N(CH <sub>3</sub> ) <sub>2</sub>	3.000	2.287	−0.713	2.351	−0.649
13*	CH <sub>2</sub> CH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.460	0.491	0.031	0.489	−0.029
14	CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.270	0.216	−0.054	0.215	−0.055
15	CH(CH <sub>3</sub> )N(CH <sub>3</sub> ) <sub>2</sub>	0.090	0.418	0.328	0.336	0.246
16*	CH(CH <sub>2</sub> CH <sub>3</sub> )N(CH <sub>3</sub> ) <sub>2</sub>	0.130	0.070	−0.060	0.003	0.127
17	CH(CH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub> )N(CH <sub>3</sub> )	0.065	0.151	0.086	0.092	0.027
18	C(CH <sub>3</sub> ) <sub>2</sub> NH <sub>2</sub>	0.145	0.449	0.304	0.287	0.142
19	C(CH <sub>3</sub> ) <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.110	0.432	0.322	0.177	0.067
20	CH <sub>2</sub> CH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.460	0.494	0.034	0.492	0.032
21*	CH(CH <sub>2</sub> CH <sub>3</sub> )CH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.400	0.393	−0.007	0.397	0.003
22	C(CH <sub>3</sub> ) <sub>2</sub> CH <sub>2</sub> NH <sub>2</sub>	0.260	0.613	0.353	0.180	−0.080
23	C(CH <sub>3</sub> ) <sub>2</sub> CH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.200	0.391	0.191	0.402	0.202
24	CHFCH <sub>2</sub> N(CH <sub>3</sub> ) <sub>2</sub>	0.590	0.625	0.035	0.658	0.068
25*		0.046	0.231	0.185	0.272	−0.226
26		0.044	0.276	0.232	0.310	0.266
27		0.120	0.087	−0.033	0.102	−0.018
28*		0.044	−0.012	−0.056	−0.010	0.054
29		0.065	−0.020	−0.085	−0.010	−0.075
30		0.120	−0.092	−0.212	−0.041	−0.161

Table 1 (continued)

Compound	R	Exp.	HM		GEP	
			Pred.	Resid.	Pred.	Resid.
31		0.095	0.178	0.083	0.199	0.104
32*		0.130	−0.032	−0.162	0.001	0.129
33		0.110	0.266	0.156	0.296	0.186
34		0.120	−0.014	−0.134	0.022	−0.098
35*		0.130	0.131	0.001	0.146	−0.016
36		0.110	0.455	0.345	0.516	0.406
37		0.070	−0.028	−0.098	0.019	−0.051
38*		0.230	0.861	0.631	1.396	−1.166

The star "\*" is test set.

### 2.3. The results of the GEP compare with the HM

In the GEP method, the  $R^2$  and  $S^2$  are 0.92 and 0.06 in training set. At the same time, in the HM method, the  $R^2$  and  $S^2$  are 0.85 and 0.11 in whole data set. The coefficient correlation of the GEP is higher than that of the HM, this is in agreement with our previous work [14]. At the same time, mean error of the GEP is lower than that of the HM.

### 3. Conclusion

QSAR approach was applied successfully to a series of 38 compounds with well-expressed anticoccidial activity. A good QSAR model with four theoretical molecular descriptors is obtained. An external validation is performed aiming to evaluate the predictive ability of the model. All descriptors involved are calculated solely from the chemical structure of the compounds and have definite biochemical meaning corresponding to the nature of the anticoccidial drugs' action. The distance of C atom with H and N, N atom number play a key role in the anticoccidial  $IC_{50}$ .

### 4. Experimental section

#### 4.1. Data preparation

The experimental values for the  $IC_{50}$  of the imidazopyridine anticoccidial compounds are taken from the literature [15] and presented in Table 1. The data set is randomly separated into a training set of 26 compounds and a test set of 12 compounds. The

training set is used to build the model and the test set is employed to evaluate the prediction ability of the model.

#### 4.2. Calculation of the descriptors

To obtain a QSAR model, compounds are often represented by the molecular descriptors. The numerical representation (often called molecular descriptor) of the chemical structure is the most important factor affecting the quality of the QSAR model. In the present investigation, the calculation process of the molecular descriptors is described as below: all molecules are drawn into HyperChem [16] and pre-optimized using MM+ molecular mechanics force field. A more precise optimization has been done with semi-empirical AM1 method in MOPAC [17]. The molecular structures are optimized using the Polak–Ribiere algorithm until the root mean square gradient is 0.01. The MOPAC output files are used by the CODESSA program [13] 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 [13]. CODESSA combines diverse methods for quantifying the

**Table 2**

Descriptors, coefficients, standard error, and *t*-values for the linear model based on the training set.

Descriptor	Chemical meaning	Coefficient	<i>t</i> -Test
Intercept		1.56E + 02	−11.162
MCICH	Min Coulombic interaction for a C–H bond	3.48E + 01	10.013
NNA	Number of N atoms	8.41E − 01	4.652
MCICN	Min Coulombic interaction for a C–N bond	1.69E + 00	5.350
MBCMO	Max bonding contribution of MO	1.62E + 01	3.120

structural information about the molecule with advanced statistical analysis to establish molecular structure–property/activity relationships. CODESSA has been applied successfully in a variety of QSAR analyses [18,19].

#### 4.3. Development of linear model by the HM [18,19]

Once the molecular descriptors are generated, the HM in CODESSA is used to pre-select the descriptors and build the linear model. The advantages of the HM are the high speed and no software restrictions on the size of the data set. The HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. The details of selecting descriptors are as follows: First of all, all descriptors are checked to ensure that values of each descriptor are available for each structure. Descriptors for which values are not available for every structure in the data 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 the insignificant descriptors are 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. The details of validating intercorrelation are (a) all quasi-orthogonal pairs of structural descriptors are selected from the initial set. Two descriptors are considered orthogonal if their intercorrelation coefficient  $r_{ij}$  is lower than 0.1; (b) CODESSA uses the pairs of orthogonal descriptors to compute the bi-parametric regression equations; (c) to an MLR model containing  $n$  descriptors, a new descriptor is added to generate a model with  $n + 1$  descriptors if the new descriptor is not significantly correlated with the previous  $n$  descriptors; step (c) is repeated until MLR models with a prescribed number of descriptors are obtained. The goodness of the correlation is tested by the square of coefficient regression ( $R^2$ ), square of cross-validate coefficient regression ( $R_{cv}^2$ ), the *F*-test (*F*), and the standard deviation ( $S^2$ ). From the above processes, five descriptors are selected from descriptors' pool and the linear model is produced by the HM. The heuristic method usually produces correlations 2–5 times faster than other methods with comparable quality [20–22].

In this equation  $b_0$  is the regression coefficient for the intercept and the  $b_i$  values are the regression coefficients (for variables 1 through  $p$ ) computed from the data.

**Table 3**

Correlation matrix of the descriptors in the model.

	MCICH	NNA	MCICN	MBCMO
MCICH	1.000	0.148	0.080	0.120
NNA		1.000	−0.020	0.070
MCICN			1.000	−0.202
MBCMO				1.000

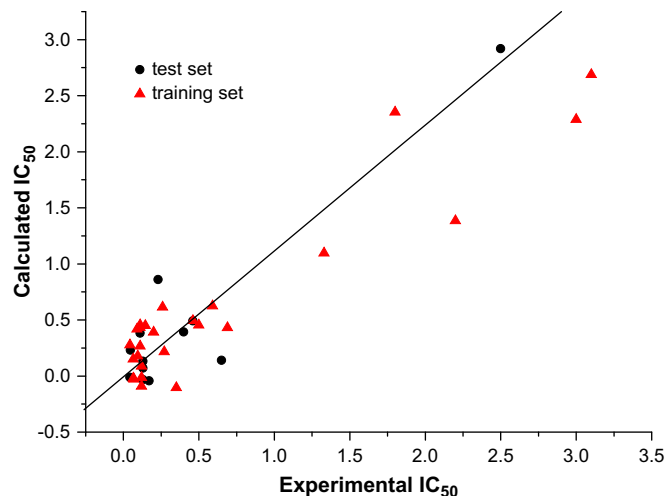


Fig. 2. Plot of experimental and calculated  $IC_{50}$  by HM.

The standard deviation in HM and PLS can be expressed as follows:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

where  $n$  is the number of compounds,  $x_i$  is the prediction value and  $\bar{x}$  is the average value.

However, the factors influencing the PADA of compounds are complex and not all of them are in linear correlation with the PADA. To develop more accurate models, nonlinear methods with the SVM and the GEP are also used.

#### 4.4. 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 [23,24]. The first step is to choose the fitting function. Mathematically, the fitness of an individual program is expressed by the equation

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

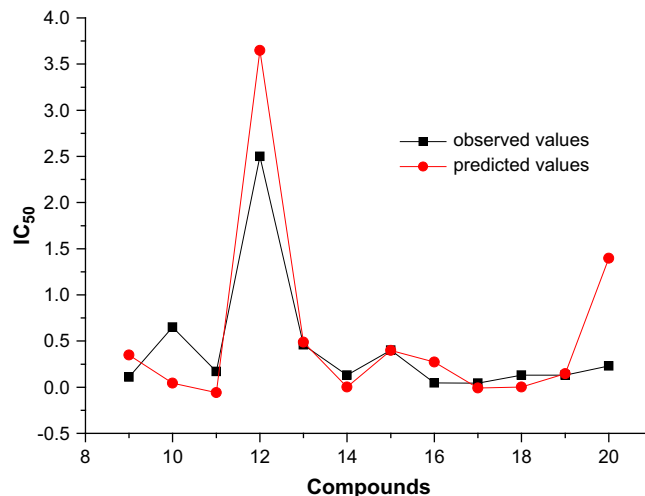


Fig. 3. Fitting curve of experimental and calculated  $IC_{50}$  in test set by GEP.

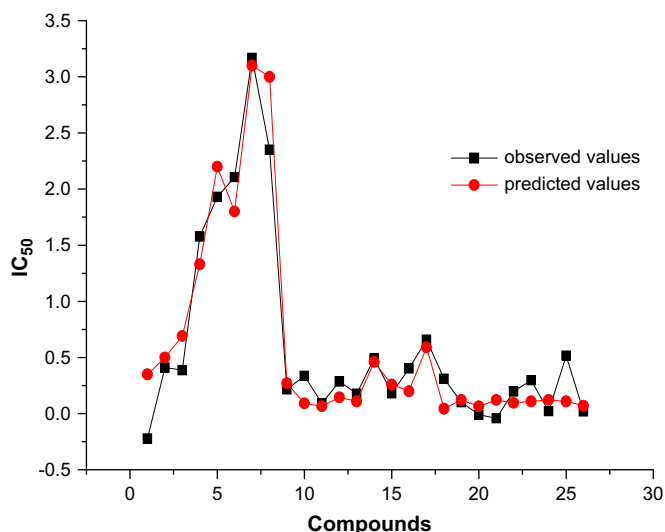


Fig. 4. Fitting curve of experimental and calculated  $IC_{50}$  in training set by GEP.

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

$$f(i,j) = \begin{cases} 1, & E_{(ij)} \leq p \\ 0, & \end{cases}$$

where  $p$  is the precision and  $E_{(ij)}$  is the relative error of an individual program  $i$  for the fitting case  $j$ . The  $E_{(ij)}$  [22] is given by

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

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 [24], 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

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

where  $\text{Cov}(T,P)$  is the covariance of the target and model outputs.  $\sigma_t$  and  $\sigma_p$  are the corresponding standard deviations.

#### 4.5. 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 1 G RAM system.

#### Acknowledgements

The authors thank the Gepsoft Team for providing the gepsoft software. We also thank the National Natural Science Foundation of China (20773071, 20703027) for financial supports.

#### References

- [1] L.R. McDougald, W.M. Reid, Coccidiosis, in: B.W. Calnek (Ed.), Diseases of Poultry, 10th ed. Iowa State University Press, 1997, pp. 865–883.
- [2] P.C. Allen, R.H. Fetterer, Clin. Microbiol. Rev. 15 (2002) 58–65.
- [3] R.G.K. Donald, J. Allocco, S.B. Singh, B. Nare, S.P. Salowe, J. Wiltsie, P.A. Liberator, Eukaryot. Cell 1 (2002) 317–328.
- [4] A.K. Bandyopadhyaya, J. Johnsamuel, A.S. Al-Madhoun, S. Eriksson, W. Tjarks, Bioorg. Med. Chem. 13 (2005) 1681–1689.
- [5] B. Gopalakrishnan, A. Khandelwal, S.A. Rajjak, N. Selvakumar, J. Das, S. Trehan, J. Iqbal, M.S. Kumar, Bioorg. Med. Chem. 11 (2003) 2569–2574.
- [6] R.G. Karki, V.M. Kulkarni, Bioorg. Med. Chem. 9 (2001) 3153–3160.
- [7] Y. Marrero-Ponce, J.A. Castillo-Garit, E. Olazabal, H.S. Serrano, A. Morales, N. Castanedo, F. Ibarra-Velarde, A. Huesca-Guillen, A.M. Sa'nchez, F. Torrens, E.A. Castro, Bioorg. Med. Chem. 13 (2005) 1005–1020.
- [8] Y. Marrero-Ponce, M. Iyarreta-Veitia, A. Montero-Torres, C. Romero-Zaldivar, C.A. Brandt, P.E. Avila, K. Kirchgatter, Y. Machado, J. Chem. Inf. Model. 45 (2005) 1082–1100.
- [9] A. Meneses-Marcel, Y. Marrero-Ponce, Y. Machado-Tugores, A. Montero-Torres, D.M. Pereira, J.A. Escario, J.J. Nogal-Ruiz, C. Ochoa, V.J. Aran, A.R. Marti'nez-Ferna'ndez, R.N. Garc'a-Sa'nchez, Bioorg. Med. Chem. Lett. 15 (2005) 3838–3843.
- [10] C. Ferreira, Adv. Complex Syst. 5 (2002) 389–408.
- [11] Ö. Terzi, M. Erol Keskin, J. Appl. Sci. 5 (2005) 508–512.
- [12] A. Baykasoglu, T. Dereli, S. Tanis, Cem. Concr. Res. 34 (2004) 2083–2090.
- [13] A.R. Katritzky, V.S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis Reference Manual, Version 2.0 (1994).
- [14] 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.
- [15] A. Scribner, R. Dennis, S. Lee, G. Ouvre, D. Perrey, M. Fisher, M. Wyvratt, P. Leavitt, P. Liberator, A. Gurnett, C. Brown, J. Mathew, D. Thompson, D. Schmatz, T. Biftu, Eur. J. Med. Chem. (2007) 1–29.
- [16] HyperChem 4.0, Hypercube, Inc., 1994.
- [17] J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, QCPE No. 455, Indiana University, Bloomington, IN, 1989.
- [18] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. 24 (1995) 279–287.

Table 4

Parameters for the simple symbolic regression problem.

Parameter names	Values
Number of generations	1000
Number of fitting cases	38
Function set	+, −, *, /, 'log', 'sqrt'
Gene head size	8
Number of genes	10
Linking function	+
Mutation rate	0.044
1-Point recombination rate	0.3
2-Point recombination rate	0.3
Gene recombination rate	0.1
IS transposition rate	0.1
IS elements' length	1, 2, 3
RIS transposition rate	0.1
RIS elements' length	1, 2, 3
Gene transposition rate	0.1
Selection range	100
Precision	0.01

- [19] H.X. Liu, C.X. Xue, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci. 44 (2004) 161–167.
- [20] H.Z. Si, T. Wang, K.J. Zhang, Z.D. Hu, B.T. Fan, Bioorg. Med. Chem. 14 (2006) 4834–4841.
- [21] A.R. Katritzky, J. Chem. Inf. Comput. Sci. 41 (2001) 1521–1530.
- [22] C.X. Xue, H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chromatogr. A 1048 (2004) 233–243.
- [23] C. Ferreira, Complex Syst. 13 (2001) 87–129.
- [24] C. Ferreira, Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence, second ed. Springer-Verlag, Germany, 2006.