Journal of Computational Science & Engineering 22 (2016) 690-697

Available at http://www.asocse.org ISSN 1710-4068

QSAR Study of 2,4-diaminopyrimidine Derivatives as GHS-R Antagonists by HM and GEP

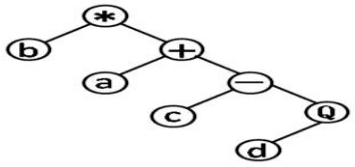
Fucheng Song[1], Lianhua Cui[1], Jinmei Piao[1], Hongzong Si[2]*, Honglin Zhai[3]

[1]School of Public Health, Qingdao University, Qingdao, Shandong 266071, China

[2]Institute for Computational Science and Engineering, Laboratory of New Fibrous Materials and Modern Textile, The Growing Base for State Key Laboratory, Qingdao University, Qingdao, Shandong 266071, China

[3]Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, China

Abstract: **Two-dimensional quantitative structure–activity relationship studies were performed on a series of 2,4-diaminopyrimidine derivatives as the inhibitors of growth hormone secretagogue receptor, and to gain insights into the structural determinants. The heuristic method was used to explore the descriptors responsible for bioactivity and gain a best linear model with R^2 0.74 in CODESSA software. Gene expression programming method produced a nonlinear quantitative model that R^2 0.80 for training set, and R^2 0.72 for test set. This paper provides new instruction for effective drug designing and screening in the future.**



Keywords: QSAR; Gene expression programming; Heuristic method; 2,4-diaminopyrimidine derivatives; GHS-R

Received 5 January 2016, Revised 23 February 2016, Accepted 25 February 2016

*Corresponding Author: Hongzong Si, sihz03@126.com

1. Introduction

Obesity has become one of the most serious public health challenges in the 21st century. Around the world, 39% of adults were overweight and 13% of all adults were obese in 2014. Nigatu [1] proposed that obesity prevalence has tripled in many countries of European Region since the 1980s. In US 16.9% of 2-19-year-olds currently meet obesity status [2]. According to China's official newspaper People's Daily, the population of obese people in China will go beyond 200 million, by 2015. According to The Brazilian Survey on Family Budgets, 12.5% of men and 16.9% of women were obese [3]. Obesity as the risk factor has an intimate relationship with great variety of diseases, such as atherosclerosis, type-2 diabetes, nonalcoholic fatty liver disease, and other immune-related disorders such as asthma and infection [4,5]. One recent analysis shows that obese people are usually rated as less happy, less likely to succeed in a relationship, less self-confident, and more likely to sink into depression [6,7].

Obesity can be diagnosed by using body mass index (BMI), one of the most commonly methods of evaluating the nutritional status, a formula based on weight in relation to height. Obesity is defined as a BMI 30kg/m^2 or higher [8-10]. The cause of obesity is complicated, and involves multifactorial factors in its development. The high intake of energy-dense food and the physical inactivity are the main factors that contribute to excessive body weight or obesity. Marieke established that some obese people are also related to increased specific eating psychopathology, decreased self-esteem, and depressed mood. The prevention and treatment of obesity include a healthy lifestyle with regular physical exercise, balanced diet, and clinical therapy. However, diet and exercise may not help obese patients achieve satisfactory results. As other non-pharmacological measures fail, the use of drugs should be considered [11]. Ghrelin has attracted tremendous attention since its discovery as an endogenous ligand for the growth hormone secretagogue receptor (GHS-R), which is also involved in the short and long-term regulation of energy balance [12,13].

Thus, GHS-R antagonist could possess the potential novel curative effect in obesity therapy. It is reported that propose that 2,4-diaminopyrimidine derivatives as potent and selective human bioavailable small molecule GHS-R antagonists could efficiently decrease food intake and body weight in several rat studies. The quantitative structure-activity relationship (QSAR) method is based on the chemical structure information and using quantitative parameters alone or combining with quantitative conventional parameters to construct the mathematical model [14], and using this model to forecast the IC50 of 2, 4-diaminopyrimidine derivatives. To our knowledge, the QSAR study plays a key role in the screening of highly active drugs.

In the two decades, many scholars had solved

2016 © The American Computational Science Society. All rights reserved.

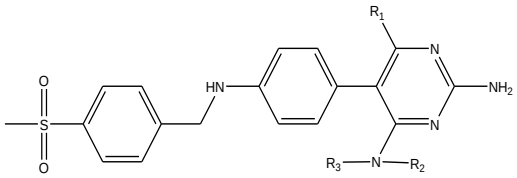
the prediction problems by the establishment of bionic mathematics calculation model and the achievements were surprising [15]. Establishing a stability and rapid classification model is what we want. Gene expression programming (GEP) is an automatic programming approach, which overcomes certain limitations of genetic algorithm and genetic programming by working with two elements, the chromosome and the expression tree successfully used in many areas, but seldom used in pharmaceutical chemistry analysis. In our work, we calculated the 54 2,4-diaminopyrimidine derivatives by the software CODESSA. Then six descriptors were selected to build several multivariable linear models by the heuristic method (HM). In order to get a better prediction model, we developed another quantitative model by GEP to predict the IC50 relationship with its molecular parameters of those GHS-R inhibitors.

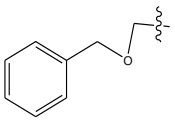
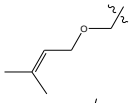
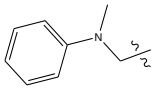
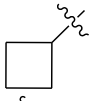
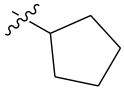
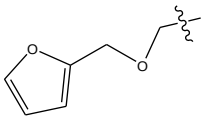
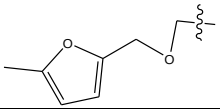
2. Data and method

2.1. Data

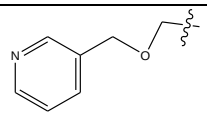
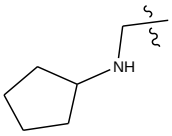
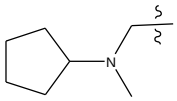
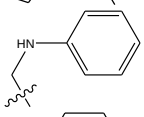
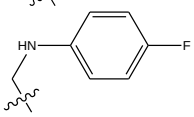
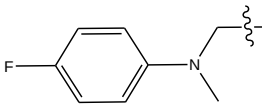
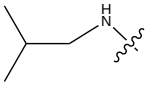
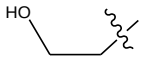
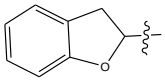
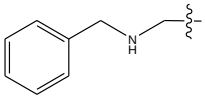
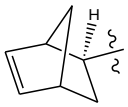
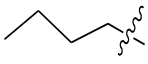
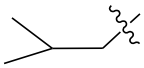
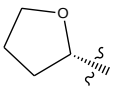
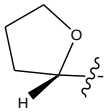
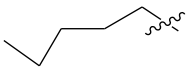
The structures and IC_{50} values for 54 2,4-diaminopyrimidine derivatives were collected from the literature [16] and are listed in Table 1. The smaller of the value, the compounds are more active. All compounds were stochastically divided into two groups: a training set of 39 compounds and a test set of 15 compounds. Values IC_{50} are averages of triplicate from FLIPR assay against human GHS-R. This research takes $LogIC_{50}$ as indicators of 2,4-diaminopyrimidine derivatives. We applied QSAR model built according to the training set to predict the corresponding results of test set. Also, the test set was used to evaluate its prediction ability.

Table 1 Experimental and predicted activities ($LogIC_{50}$) of 2,4-diaminopyrimidine derivatives

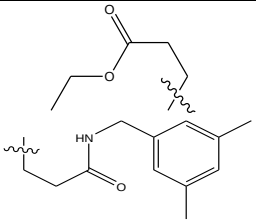
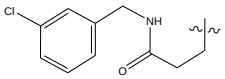
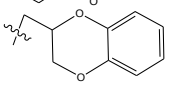
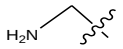
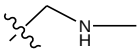
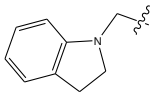
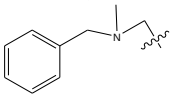
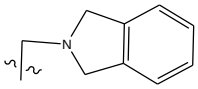
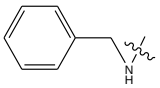
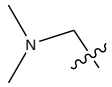
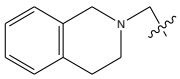
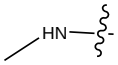
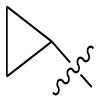
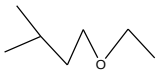
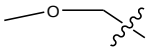
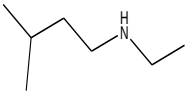
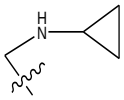
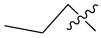


Compounds	R_1	R_2	R_3	$LogIC_{50}$	GEP		HM	
					Pre.	Resid.	Pre.	Resid.
1	Ethyl	H	H	1.57	1.66	-0.09	1.53	-0.04
2*		H	H	0.85	1.01	-0.16	0.80	-0.05
3		H	H	2.14	1.85	0.29	2.15	0.01
4*		H	H	0.28	1.04	-0.76	1.18	0.90
5		H	H	0.95	1.23	-0.28	0.96	0.01
6		H	H	1.15	1.84	-0.69	1.81	0.66
7		H	H	1.32	1.67	-0.35	1.61	0.29
8		H	H	1.56	1.51	0.05	1.37	-0.19
9		H	H	0.98	0.95	0.03	1.41	0.43
10		H	H	0.99	1.1	-0.11	0.83	-0.16

2016 © The American Computational Science Society. All rights reserved.

11*		H	H	2.27	1.47	0.8	2.05	-0.22
12*		H	H	3.54	2.27	1.27	2.54	-1.00
13*		H	H	2.50	2.65	-0.15	2.73	0.23
14		H	H	0.96	1.05	-0.09	1.04	0.08
15		H	H	1.08	1.11	-0.03	1.22	0.14
16		H	H	1.00	1.33	-0.33	1.36	0.36
17		H	H	1.15	1.46	-0.31	1.67	0.52
18*	Ethyl		H	3.39	2.41	0.98	2.38	-1.01
19	Ethyl		H	3.49	3.42	0.07	3.29	-0.20
20		H	H	1.58	1.44	0.14	1.98	0.40
21		H	H	1.69	1.8	-0.11	1.73	0.04
22*		H	H	1.82	2.05	-0.23	2.38	0.56
23		H	H	2.45	2.17	0.28	2.40	-0.05
24		H	H	1.51	1.85	-0.34	1.50	-0.01
25		H	H	1.60	1.59	0.01	1.20	-0.41
26*		H	H	1.73	1.91	-0.18	1.99	0.26
27		H	H	1.76	1.49	0.27	1.64	-0.12
28		H	H	1.04	1.53	-0.49	1.08	0.04
29		H	H	2.32	1.98	0.34	1.69	-0.63
30		H	H	1.79	1.83	-0.04	1.78	-0.01

2016 © The American Computational Science Society. All rights reserved.

31*		H	H	1.91	2.35	-0.44	2.19	0.28
32		H	H	0.48	1.02	-0.54	1.10	0.62
33		H	H	0.73	0.48	0.25	0.12	-0.61
34		H	H	2.26	1.19	1.07	1.34	-0.92
35*		H	H	3.85	2.74	1.11	3.18	-0.67
36		H	H	3.14	3.24	-0.1	3.25	0.11
37		H	H	1.11	1.3	-0.19	1.60	0.49
38		H	H	1.84	2.34	-0.5	2.16	0.32
39*		H	H	1.53	1.51	0.02	1.13	-0.40
40	Ethyl		H	2.55	2.43	0.12	1.95	-0.60
41		H	H	3.41	3.59	-0.18	4.06	0.65
42		H	H	1.61	1.47	0.14	1.57	-0.04
43*	Ethyl		H	1.84	2.01	-0.17	2.15	0.31
44	Ethyl		H	2.89	2.28	0.61	2.33	-0.56
45		H	H	1.58	1.83	-0.25	1.93	0.35
46		H	H	1.76	1.96	-0.2	2.36	0.60
47		H	H	1.28	1.17	0.11	1.50	0.22
48*		H	H	0.83	1.06	-0.23	1.15	0.32
49		H	H	1.72	1.7	0.02	1.19	-0.53
50*		H	H	3.15	2.99	0.16	2.94	-0.21
51		H	H	2.98	2.47	0.51	2.64	-0.34
52*	Ethyl	Ethyl	Ethyl	2.58	2.10	0.48	2.48	-0.10
53	Ethyl		H	2.37	2.37	0	2.52	0.15
54	Ethyl		H	2.87	2.51	0.36	2.58	-0.29

* Test set.

2016 © The American Computational Science Society. All rights reserved.

2.2. Calculation of the descriptors

In QSAR study, the molecular physical and chemical parameters are generally adopted as the numerical descriptors of compounds. These descriptors of chemical information are used to form a complex mathematical formula that can reflect the relationship between biological activity and structure parameters of compounds. All molecules structures were drawn into Hyperchem, the compounds were not only pre-optimized utilizing MM+ molecular mechanics force field, but also got a more precise optimization by the semi-empirical AM1 method in MOPAC. The Semi-empirical AM1 method can get more accurate optimization. We used the Polak–Ribiere algorithm to optimize the molecular structures until the root mean square gradient reaches 0.01. Then, the HIN files generated by geometrical optimization, the MNO files generated by MOPAC calculation and the IC50 values will be imported to CODESSA software to select the optimum molecular descriptors from six groups of descriptors: Constitutional descriptors, Geometrical descriptors, Electrostatic descriptors, Thermodynamic descriptors, CPSA descriptors and Topological descriptors.

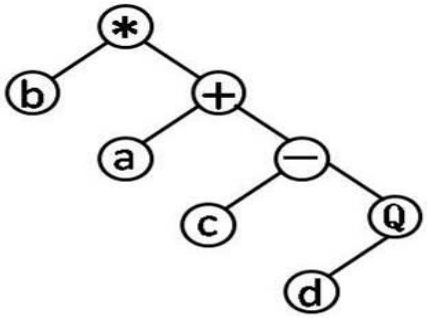


Figure 1. Expression trees.

2.3. Heuristic method

Heuristic method (HM) could search for a large number of molecular descriptors to create the best linear equation in CODESSA software. First of all, the collinearity of molecular descriptors is in control. If the correlation coefficient of any two descriptors is greater than 0.8, the descriptors will not appear in the same model. Then, all descriptors are checked to ensure: (1) each descriptor is available for whole compounds and (2) values of descriptor are different from each other. Heuristic method screens rapidly Every Parameter is screened so that could build the best model, but rather than a random combination of parameters. We used the coefficient of determination (R^2), square of cross-validate coefficient regression (R^2_{CV}), F test value (F), the standard deviation (S^2) to evaluate the stability of HM model. In our study, the root-mean-square of HM can be calculated by Eq (1):

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(y_k - \hat{y}_k)}{n_s}} \quad (1)$$

y_k : the predictive value, $\hat{y}_k$: the experimental value, n_s : the number of compounds.

2.4. Gene expression programming

GEP is a novel method of automatic programming that based on the structure and function of biological gene. It was invented by scientists Ferreira of Portuguese and was developed from genetic algorithms and genetic programming. GEP has all the characteristics of GA and GP, commonly used selection operator, mutation operator, 1-point and 2-point operator. Expression tree generated successively from scanning genotype individual element according to the hierarchical order, such as Figure 1.

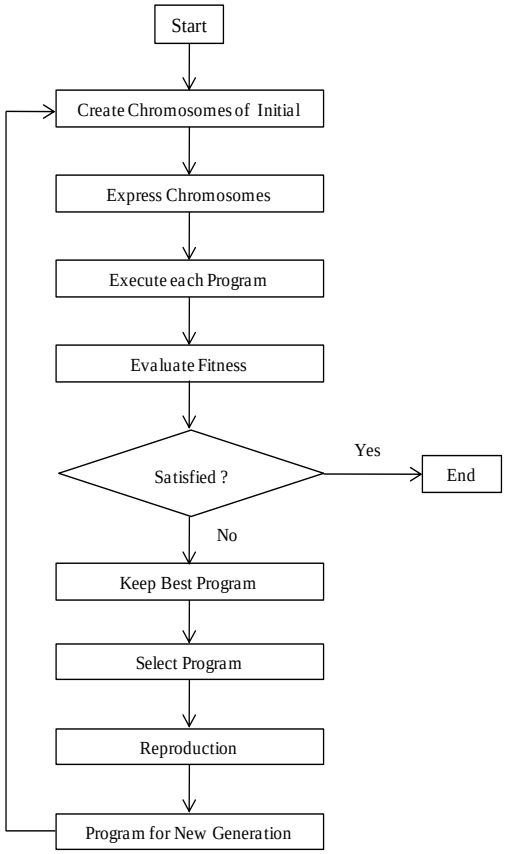


Figure 2. The flow chart of gene expression algorithm.

2016 © The American Computational Science Society. All rights reserved.

The major step is to choose the linking function for gene expression programming. The common set of functions, $F=\{+,-,*,\div,Q\}$, Q represents the root function. Expression trees in Figure 1 can express in a function, $F=b*\left(a+\left(c-\sqrt{d}\right)\right)$. The application of complex function can improve the prediction ability of QSAR model. GEP genetic operators don't only include the selection operator and mutation operator of genetic algorithm, but also increase the conversion, insert and restructuring operator. Flow chart of GEP classification algorithm is shown in Figure 2. This Process will repeat for a preestablish number of generations until what a best model has been found.

Table 2 Physicochemical meaning and statistical parameters of each descriptor

Symbol	Physical-chemical meaning	Coefficient	Standard error	T-test
N_{min}	Min partial charge for a N atom	-710.85	76.540	-9.2873
DPSA1	Difference between partial positively and negatively charged surface areas	0.0086	0.001	5.2347
RPC	Relative positive charge	59.692	11.011	5.4211
N_{max}	Max bond order of a N atom	32.957	8.374	3.9357
IOKSE	Image of the Onsager-Kirk wood solvation energy	10.895	3.283	3.3176
AVH	Avg valency of a H atom	- 281.53	97.656	-2.8829

3. Results and discussion

3.1. The HM model

373 descriptors were calculated and analyzed by the CODESSA program for all the 2,4-diaminopyrimidine compounds. The CODESSA software can give different HM equation that depending on the number of descriptors selected. HM can select a series of descriptors that are most relevant to the IC50 of the GHS-R inhibitors. We found that the values of R^2, R^2_{cv}, and S^2 changed with the increase of molecular descriptors number. The trend of 1-7 variables is shown in Figure 3. Finally, when selecting 6 descriptors, the HM model tended to be stable. And these descriptors were used to build a HM equation (2). The corresponding physicochemical meaning and statistical parameters are in the table 2.

$$LogIC_{50}= 144.21-710.85N_{min} + 0.0086DPSA1 + 59.692RPC + 32.957N_{max} + 10.895IOKSE - 281.53AVH \quad (2)$$

$R^2 = 0.74$, $R^2_{cv} = 0.65$, $S^2 = 0.219$, F= 22.0692

In order to avert the possible multi-collinearity of different descriptor scales, table 3 lists the correlation coefficients between the 6 groups descriptor tested by SPSS 20.0. Obviously, there is any correlation coefficient more than 0.8. It really shows that all the descriptors are strongly orthogonal, and the HM model has a pretty statistical reliability for predicting the IC50 of 2,4-diaminopyrimidine derivatives.

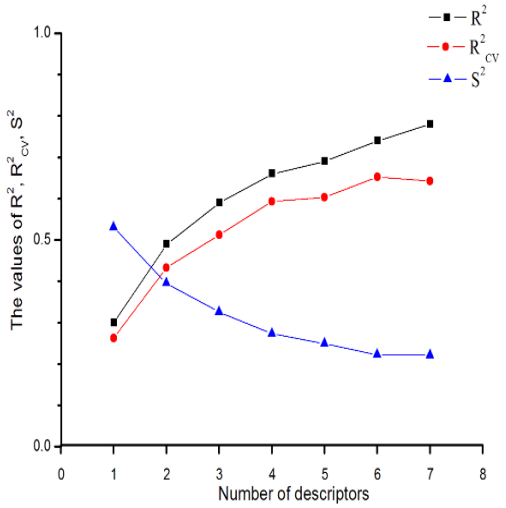


Figure 3. Different number of descriptors with the corresponding R^2, R^2_{CV}, S^2.

By explanation of the descriptors selected, it is possible to gain some intrinsic mechanisms that are likely to reveal the depletion of IC_{50} of 2,4-diaminopyrimidine derivatives. The correlation coefficients can be order as: N_{min} > AVH > RPC > N_{max} > IOKSE > DPSA1.

Table 3 Correlation coefficient of the 6 descriptors

	N_{min}	DPSA1	RPC	N_{max}	IOKSE	AVH
N_{min}	1.000	0.180	0.083	0.373	0.177	0.077
DPSA1		1.000	0.156	0.201	0.282	0.767
RPC			1.000	0.213	0.096	0.449
N_{max}				1.000	0.233	0.254
IOKSE					1.000	0.317
AVH						1.000

2016 © The American Computational Science Society. All rights reserved.

Min partial charge for a N atom (N_{min}) and Max bond order of a N atom (N_{max}) are topological descriptors. The topology parameters have become an important tool of QSAR research. We usually use the distance matrix and links matrix to edit the graphics in the process of mathematical molecules and operations. Topology parameter is the simple vector or digital of molecular structure. N_{min} exposes in all models, and it has the largest negative coefficients. Min partial charge of nitrogen atom can influence the combination of drug molecules with protein. The increase of partial charge will enhance the effect of molecules with protein receptor. As a result, the activity of the protein is inhibited. By contrast, N_{max} is positively correlated to the values of IC_{50}. Greater the resonance energy between nitrogen atom and others is, more stable of the structure is. It will unable to expose the effective groups if the bong is hard to break.

Difference between partial positively and negatively charged surface areas (DPSA1) and Relative positive charge (RPC) are CPSA Descriptors. CPSA descriptors were originally designed for use in structure-physical relationship studies to capture information about the features of molecules responsible for polar intermolecular interactions. Most recently, the CPSA descriptors are examined in more detail, evaluating their characteristics with regard to conformational dependence, sources of partial atomic charges, utility of whole molecule and substructure varieties, and the inclusion or exclusion of explicit hydrogens. Stanton [17] thinks that the CPSA descriptors have been found to be practically useful in the study of acute aquatic toxicity and discriminate agonists and antagonists among compounds. DPSA1 can be calculated by eq. (3):

DPSA1 = PPSA1 – PNSA1 (3)

PPSA1-positively charged solvent-accessible molecular surface area

PNSA1-negatively charged solvent-accessible molecular surface area

Charged surface areas hinder the molecules to get into proteins and interfere with the formation of hydrogen bonds. The same principle, the greater values of relative positive charge can increase the difficulty of combination of drug molecules with GHS-R. DPSA1 and RPC play positive effects on the value of IC_{50}. They will lead to an increase in the value of IC_{50}. DPSA1 has the smallest positive coefficients while RPC has the largest.

Image of the Onsager-Kirk wood solvation energy (IOKSE) and Avg valency of a H atom (AVH) are Quantum chemical descriptors. Quantum chemical parameters can fully describe the compound structure, and the parameters reflect the precise physical property that is easy to obtain. The application of quantum chemistry has become a major direction of QSAR analysis. Image of the Onsager-Kirk wood solvation energy refers that the binding energy between the solute and the solvent in the solution. Drug molecules with high binding energy are more likely stay within the body fluids instead of combining with receptors. So the values of IOKSE are positively correlated to the values of IC50. It is well known that, molecules are usually combined with protein by hydrogen bonds [18] The high average valency of hydrogen will promote the integration of molecules to the specific protein. It can explain why the AVH is negatively correlated with IC50.

3.2. The GEP model

GEP model adopts the same molecular descriptors with HM model. The software automatic problem solver (APS) allows the easy optimization of intermediate solutions and the easy testing of the evolved models against a test set. This article used the sum of squared residuals as a fitness function:

$$\text{Fitness} = \sum_{j=1}^{n} \left(y_j - \hat{y}_j \right) \quad (4)$$

where y_j and $\hat{y}_j$ are the real and predicted output values for the jth sample respectively; N: the number of nitrobenzene compounds.

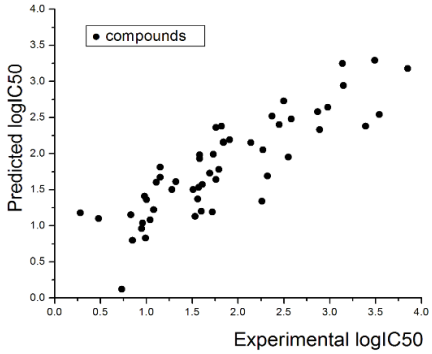


Figure 4. Plot of predicted and experimental values for HM.

In this run, made a very satisfactory solution in the 137th generation with a R^2 0.80 and S^2 0.112 for the training set and a R^2 0.72 and S^2 0.379 for the test set. Figure 5 and Figure 6 showed the fitting situation of training set and test set. The ETs can be expressed as Eq. (5)

$$\begin{aligned} LogIC50 = {} & \tan\left(\text{fmod}\left(x_2, \frac{\tan(x_5 \times (x_5 - x_4 - x_6))}{\text{pow}(10, x_1)}\right)\right) + x_6 + \\ & \text{fmod}\left(\tan\left(\text{pow}(10, x_1) + \frac{x_6}{x_3}\right), \text{pow}(10, x_5^2) \times x_1\right) + x_2 \times x_3^2 \times x_4 + \\ & \text{fmod}(x_2, \text{pow}(10, x_5 \times \text{apslogi}(2x_4 \times x_6))) + \tan x_5 + \sin x_4 + \\ & \sin\left(\text{fmod}\left(\left(\frac{\text{pow}(10, \text{apslogi}(x_2))}{x_1} + \text{fmod}((x_2 + x_6), (x_2 - x_6))\right), x_2\right)\right) \end{aligned} \quad (5)$$

In the Eq. (5) x_1, x_2, x_3, x_4, x_5, x_6 represent variables of N_{min}, DPSA1, RPC, N_{max}, IOKSE, AVH.

2016 © The American Computational Science Society. All rights reserved.

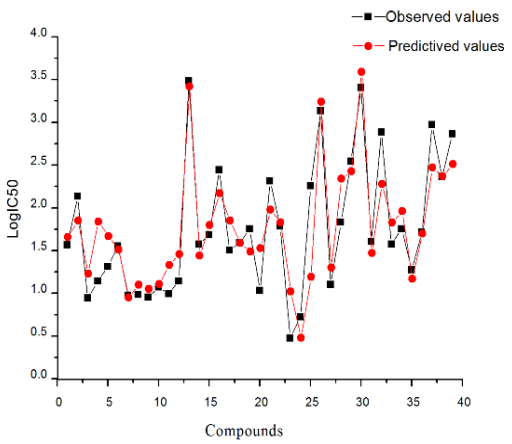


Figure 5. Fitting curve of training set.

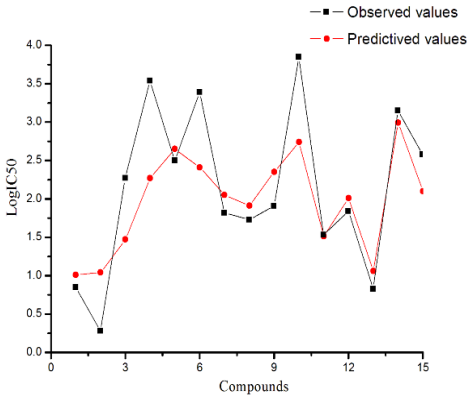


Figure 6. Fitting curve of test set.

3.3. The GEP compares with HM

In the GEP method, the R^2 and S^2 are 0.80 and 0.122 for the training set, 0.72 and 0.379 for the test set, respectively. At the same time, the R^2 and S^2 are 0.73 and 0.219 for all compounds by HM method. The results of GEP method are better than that of HM method. The GEP model has a better prediction of the $LogIC_{50}$ than that of HM.

HM only explores the linear model, but GEP can give a nonlinear model. The inhibition of drugs is so complex that a simple linear model cannot meet the needs of the QSAR analysis. At the same time, selecting molecular descriptors about physical chemistry properties is the core stage for the construction of QSAR model and a good prediction of IC_{50}.

4. Conclusion

Studies have shown that the GEP is very efficient tools to build a nonlinear model and a very effective to predict the IC_{50} of 2,4-diaminopyrimidine derivatives. The GEP has a strong ability to construct models due to the techniques of converging to the global optimum. The results show that the GEP method has better prediction ability than that of HM. It is a good approach for predicting the expected activity of drugs and aiding drug design. In the meantime, QSAR model can identify some of the internal factors that are related to the biological activity IC50 and provide some instruction for further designing of the new highly active 2,4-diaminopyrimidine inhibitors of GHS-R.

References

[1] Y.T. Nigatu, H.A. Ven, J.J. Klink, S. Brouwer, S.A. Reijneveld, U. Bültmann. Appl. Ergon., 52 (2016) 128-134.
[2] C.L. Ogden, M.D. Carroll, B.K. Kit, K.M. Flegal. J. Am. Med. Assoc., 311 (2014) 806-814.
[3] L.M. Cercato, P.A. White, F.K. Nampo, M.R. Santos, E.A. Camargo. J. Ethnopharmacol., 176 (2015) 286–296.
[4] D.P. Guh, W. Zhang, N. Bansback, Z. Amarsi, C.L. Birmingham, A.H. Anis. BMC. Public. Health., 9 (2009) 88.
[5] S.J. Kong, J.Q. Hou, M. Xia, Y. Yang, A.L. Xu, (2014). Can. Cell. Res., 1 (2014): 37-41.
[6] R.C. Shelton, M. Falola, L. Li, J. Zajecka, M. Fava, G.I., Papakostas. J. Psychiat. Res. 70 (2015) 91-97.
[7] M.Q. Werrij, S. Mulkens, H.J. Hospers, A. Jansen. Patient. Educ. Couns., 62 (2006) 126–131.
[8] Y. Seino, D. Yabe. J. Diabetes. Investig. 4 (2013) 108–130.
[9] J. Seufert, B. Gallwitz. TDiabetes. Obes. Metab., 16 (2014) 673–88.
[10] A.H. Mello, M. Prá, L.C. Cardoso, R. de Bona Schraiber, G.T. Rezin. Diabetes. Obes. Metab., 64 (2015) 967-981.
[11] W. Coutinho. Arq. Bras. Endocrinol. Metabol., 53 (2009) 262–270.
[12] D.E. Cummings, J. Overduin, M.H. Shannon, K.E. Foster-Schubert. Surg. Obes. Relat. Dis., 1 (2005) 358–368.
[13] H. Ueno, H. Yamaguchi, K. Kangawa, M. Nakazato. Regul. Pept., 126 (2005) 11-19.
[14] J. Du, H.X. Liu, X.J. Yao. J. Comput. Sci. Eng., 13 (2014) 471-479.
[15] M. Uysala, H. Tanyildizi. Constr. Build. Mater., 27 (2012) 404–414.
[16] B. Liu, M. Liu, Z. Xin, H. Zhao, M.D. Serby, C. Kosogof, L.T. Nelson, B.G. Szczepankiewicz, W. Kaszubska, V.G. Schaefer, H.D. Falls, C.W. Lin, C.A. Collins, H.L. Sham, G. Liu. Med. Chem. Lett., 16 (2006) 1864–1868.
[17] D.T. Stanton, S. Dimitrov, V. Grancharov, O.G. Mekenyan. Sar. Qsar. Environ. Res., 13 (2002) 341-351.
[18] T. Ueno, S. Abe, T. Koshiyama, T. Ohki, T. Hikage, Y. Watanabe. Chem., 16(2010): 2730-2740.

2016 © The American Computational Science Society. All rights reserved.