Journal of Computational Science &Engineering 48 (2020) 1146-1156

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

QSAR models on aminopyrazole-substituted resorcylate compounds as Hsp90 inhibitors

Xingyuan Wang[1], Haijie Dong[2*], Qingao Qin[1]

[1] School of Data Science and Software Engineering, Qingdao University, 266071, China

[2] Institute for Computational Science Engineering and Software Engineering, Qingdao University, 266071, China

Abstract: **The new Hsp90 inhibitors, aminopyrazole-substituted resorcylate compounds, has extensive and effective Hsp90 inhibitory activity and was designed to develop new antibacterial drugs. Quantitative structure-activity relationship method was used to predict fungal selectivity of new Hsp90 inhibitors. The correlation coefficient R of the best linear model was 0.89. Two nonlinear models were established by gene expression programming and back propagation artificial neural network. The correlation coefficient and mean square error of the train set of gene expression programming model were 0.88 and 0.10, the test set were 0.89 and 0.11 respectively. In the back propagation artificial neural network model, the correlation coefficient of the train set was 0.95, the correlation coefficient of the test set was 0.95, the correlation coefficient of the verification set was 0.95, and the mean square error was 0.0069. The best satisfied model will provide reference for the design and development of new Hsp90 inhibitors.**

Keywords: Hsp90 Inhibitors; Quantitative structure activity relationship (QSAR); Gene expression programming (GEP); Back propagation artificial neural network (BPANN); Heuristic method

Received 26 October 2020, Revised 22 November 2020, Accepted 28 December 2020

1.Introduction

In recent years, the incidence of clinical fungal infection has been on the rise. More than a billion people are infected with superficial infections such as ringworm and athlete's foot. Of most concern is the growing number of invasive systemic infections, which kill more than a million people a year [1]. According to the international standard, fungal infections can be divided into three categories, namely superficial mycosis, dermatomycosis and systemic mycosis [2-3]. More than 90% of the pathogens of invasive mycosis are Candida albicans, Aspergillus fumigatus, and Cryptococcus neoformans. In recent years, due to the abuse of antibiotics and other problems lead to the emergence of drug-resistant strains and superbugs, the existing antifungal drugs are increasingly unable to meet the clinical needs [4]. In fact, the only new class of antifungals to reach the clinic in decades has no efficacy against C. neoformans and related species [5].

Hsp90 is a highly conserved molecular chaperone that is ubiquitous in both eukaryotic and prokaryotic cells. The main function of Hsp90 is to bind receptor proteins with tertiary structure and participate in the activation and maturation of proteins. Hsp90 forms a complex with other helper molecular chaperones and binds to many "client proteins" in the cell, facilitating the proper folding and assembly of these proteins to achieve their active conformation and enhance their stability. Thereby allowing them to be transported through different channels. Hsp90 has been extensively validated as a regulator of virulence and antifungal drug resistance in Candida and Aspergillus species [6-7]. By changing the conformation of steroid hormone receptor, Hsp90 molecular chaperone complex can regulate the signal transduction pathway of cells, improve the adaptability of cells to external stimuli, enhance the survival ability of microbial body cells, and enhance the tolerance to antimicrobial drugs [8]. The study of Cowen et al. [9] showed that the higher the intracellular Hsp90 content, the stronger the fungal resistance, which requires the presence of Hsp90 for the generation of resistance. If Hsp90 is absent, the resistance will also disappear. Therefore, currently, Hsp90 inhibitors are an important method to treat microbial resistance. And fungal selectivity is a key characteristic of the development of Hsp90 inhibitors as antifungal agents. In recent years, the abuse of antibiotics and other problems lead to the emergence of drug-resistant strains, the existing antifungal drugs are increasingly unable to meet the clinical needs, while the development of new antifungal drugs such as Hsp90 inhibitors often requires a lot of time and money [10]. The quantitative structure-activity relationship (QSAR) method can be used to design the synthesis of new drugs by predicting the fungal selectivity of new Hsp90 inhibitors.

QSAR is a science of studying the mathematical relationship between structure and properties by mathematical and statistical methods based on the structure of matter molecules. It can present the quantitative change law between the structure and biological activity or physicochemical property, and use this law to predict the biological activity or property of new compound, and reveal the structural factors that affect the activity. It is one of the most effective methods to optimize lead compounds and design new drugs[11]. QSAR attempts to find a consistent relationship between biological activity and molecular properties, providing useful information for the

2020@ The American Computational Science Society. All rights reserved.

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

design of new compounds acting on specific targets. An excellent QSAR model helps to understand the mechanism of drug action and can save the cost and time of new drug development [12-14]. However, there are many commonly used methods for establishing the model. Among them, the methods for establishing the linear model include heuristic method (HM), multiple linear regression (MLR), partial least square (PLS), etc. The methods for establishing the non-linear model include back propagation artificial neural network (BP-ANN), support vector machine (SVM), gene expression programming (GEP) etc. These methods are widely used in biological medicine, chemical engineering and other scientific research. Heuristic methods speed up the process of reaching a satisfactory solution, and molecular descriptors can be calculated in a short time. MLR is the most common statistical method in the study of two-dimensional structure-activity relationship. PLS considers both independent and dependent variables, so that the model can describe both independent and dependent variables. Compared with MLR, the prediction residual sum of squares (PRESS) of PLS model is smaller, so it has higher prediction stability. SVM is a supervised learning method. It can be widely used in statistical classification and regression analysis by kernel method for nonlinear classification [15]. GEP, as a relatively new genetic algorithm, has attracted more and more attention in academic circles for its excellent characteristics in formula mining, symbol regression, etc [16]. BP-ANN is the most widely used neural network algorithm, which has been proved to have the function of realizing any complex nonlinear mapping and has strong self-learning ability. In this study, BP-ANN was selected to construct high predicting model for the future compounds selection.

2. Material and methods

2.1. Data set

89 aminopyrazole-substituted resorcylate compounds and the half maximal effective concentration (EC50) were collected from the literature[17] and were listed in Table 1. To obtain better results, the conversion of the EC50 value was converted to a negative logarithmic scale, i.e. pEC50. It is a common QSAR analysis indicator. Compared with half of the maximum effective concentration, the higher the pEC50 value, the greater the fungal selectivity, and vice versa. In order to make the model more accurate, 89 inhibitors were randomly divided into train set and test set according to the ratio of 4:1. The train set contained 71 small molecule inhibitors and the test set contained 18 small molecule inhibitors.

2.2. Calculation of the descriptors

Molecular descriptor refers to the measurement of a molecule's properties in a certain aspect, which can be either the physical and chemical properties of the molecule or the numerical index pushed out by the molecule through various algorithms[18].In this study, in the first step, the molecular structures of 89 amino pyrazole substituted isophthalamide compounds were drawn in ChemDraw software. In the second step, the drawn molecular structures were imported into Hyperchem software one by one, and the MM+ molecular mechanics force field was used for rough optimization. Then the semi-empirical AM1 or PM3 method was used in HyperChem for precise optimization, and the Polak-Ribiere algorithm was used to optimize the molecular structure until the root mean square gradient reached 0.01. Saved MOPAC output file as two copies with different suffix names, and added pEC50 value respectively. These two files were eventually used to calculate 7 molecular descriptors in CODESSA software.

2.3. Linear model by HM[19]

After all the molecular descriptors were calculated, these molecular descriptors needed to be screened. Feature selection was to reduce the number of descriptors and remove relatively unimportant descriptors, while the remaining descriptors could fully represent the molecular structure and physical and chemical properties as much as possible. Heuristic method (HM) in CODESSA software can search a large number of descriptors completely, and the best regression equation can be established without examining all possible parameter combinations. HM mainly includes the following four steps: ①Selection of 1-parameter descriptor. The square of correlation coefficient (R^2), F-test and t-test are used as selection criteria, and descriptors with low correlation with properties(activity) are removed. Remove descriptors without significant changes; Remove some descriptors with high correlation. ②Selection of 2-Parameter Descriptor. R^2 and F- tests are the criteria for analysis and selection. ③Selection of n-parameter descriptor. After obtaining the two-parameter correlation coefficients with optimal statistical characteristics, the descriptors not used in the previous selection process are added to establish a new correlation equation. The new correlation coefficient is verified by R^2, F-test and standard deviation(S). Until the established correlation equation contains the maximum number of parameters. The process of adding descriptors starts with the correlation equation with the maximum fitness function value, and the fitness function is defined as follows:

$$w = \left(R^2 * F * n\right) / \left(N * s^2\right)$$

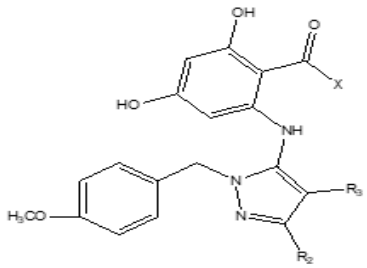
R^2 is the square of the correlation coefficient; F is F-test value; n is the number of samples; N is the number of descriptors; S is the standard deviation.

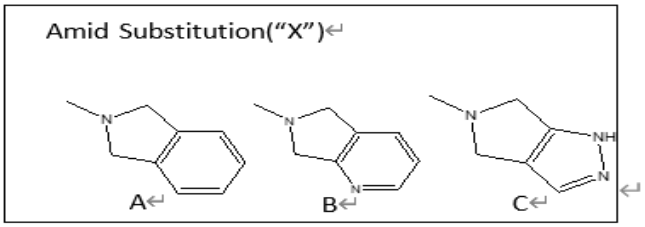
2020@ The American Computational Science Society. All rights reserved.

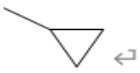
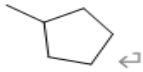
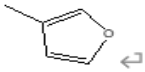
Available at http://www.asocse.org ISSN 1710-4068

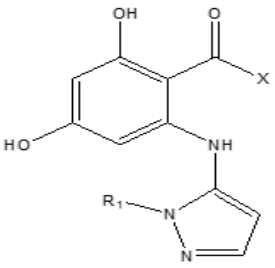
This process is repeated until the number of descriptors required by the user is reached. ④Output. Every time a descriptor is added, the optimal 10 correlation results are displayed, and every iteration starts with the results with the best correlation. The correlation between descriptors and the square of the cross-validation coefficien(R^2_{CV}) should be calculated during each cycle.

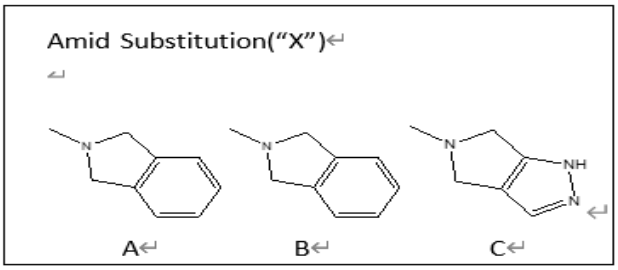
Table 1. aminopyrazole-substituted resorcylate compounds and their half maximal effective concentration (EC50)





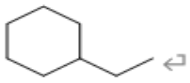
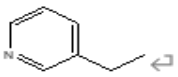
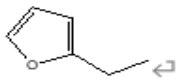
Compound	X	R_2	R_3	EC_{50}
1	A		^{i}Pr	0.396
2	A	H	Ph	1.623
3	A		Bn	1.756
4	B	Et	H	0.022
5	C		H	0.025
6[T]	B	^{i}Pr	H	0.025
7	C		H	0.023
8[V]	A	^{t}Bu	H	1.026
9[T]	B		H	0.026
10	C		H	0.023
11	B		H	0.021
12[T]	C		H	0.041
13	B		H	0.039
14	C		H	0.040





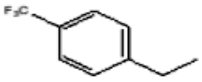
2020@ The American Computational Science Society. All rights reserved.

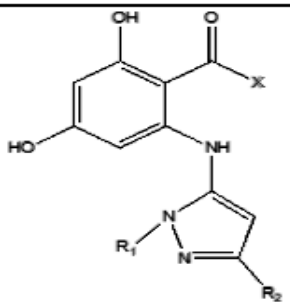
Available at http://www.asocse.org ISSN 1710-4068

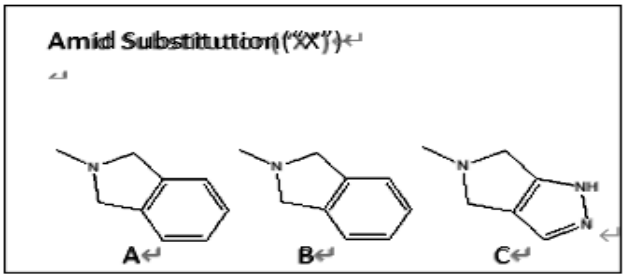
Compound	X	R_1	EC_{50}
15[V]	A		0.088
16	B	CH_3	0.286
17[T]	C		0.142
18	A		0.125
19[T]	B	^{i}Pr	0.250
20	C		0.097
21	A		0.062
22[V]	B	^{i}Bu	0.103
23	A		0.051
24[T]	B		0.026
25	C		0.018
26	A		0.061
27	B	Ph	0.044
28	C		0.037
29[T]	A		0.087
30	B	Cy	0.045
31[V]	C		0.035
32	A		0.052
33	B	Bn	0.043
34	C		0.036
35	A		0.058
36[V]	B		0.207
37	C		0.199
38[T]	A		0.074
39	B		0.106
40	C		0.058
41	A		0.044
42[T]	B		0.040
43[V]	C		0.036

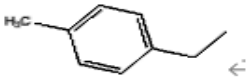
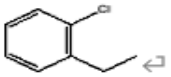
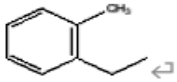
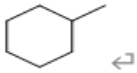
2020@ The American Computational Science Society. All rights reserved.

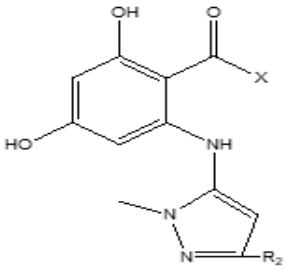
Available at http://www.asocse.org ISSN 1710-4068

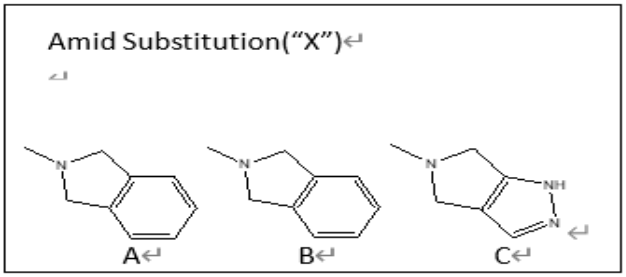
44	A		0.063
45	B		0.048





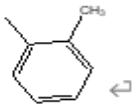
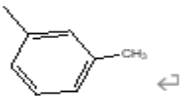
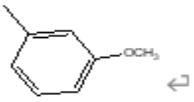
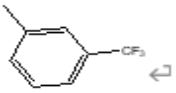
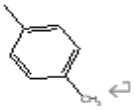
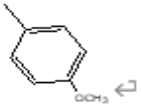
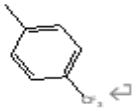
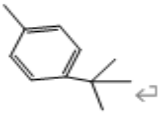
Compounds	X	R_1	R_2	EC_{50}
46[T]	A		CH_3	0.252
47	A		CH_3	0.560
48	A		CH_3	0.700
49	A		Ph	0.078
50[T]	B	CH_3	Ph	0.127
51	C		Ph	0.066
52	B		Ph	0.517
		^{t}Bu		
53	C		Ph	0.379
54[V]	B		Ph	0.615
55	C		Ph	0.419
56	B		Ph	0.315
		^{i}Bu		
57	C		Ph	0.147





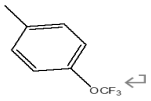
2020@ The American Computational Science Society. All rights reserved.

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

Compound	X	R_2	EC_{50}
58	A		0.087
59	B	^{i}Pr	0.062
60[T]	C		0.040
61	A		0.156
62	B	Cy	0.110
63[T]	C		0.037
64	A		0.065
65	B		0.139
66[T]	C		0.084
67	A		0.267
68[V]	B		0.421
69	C		0.176
70	A		0.281
71	B		0.852
72	C		0.244
73	A		4.601
74	B		1.523
75	C		0.434
76	A		1.318
77[V]	B		1.781
78[T]	C		0.424
79	A		0.630
80	B		1.289
81[T]	C		0.489
82[T]	A		9.530
83[T]	B		5.815
84	C		1.385
85	B		8.337
86[V]	C		2.262

2020@ The American Computational Science Society. All rights reserved.

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

87	A		5.402
88	B		1.262
89	C		0.514

T is the test set of GEP and BP-ANN, V is the verification set of BP-ANN

2.4. Nonlinear model by GEP

At the end of the 20th century, Ferreira proposed a new evolutionary algorithm based on genotype and phenotype: Gene Expression Programming (GEP). The fundamental difference between GEP and GA and GP is that they adopt different individual properties. In GA, individuals are linear strings (chromosomes) of fixed length, and in GP, individuals are nonlinear entities with different lengths and shapes. In GEP, individuals are encoded into linear strings (genomes or chromosomes) of fixed length, and then expressed into nonlinear entities of different lengths and shapes (simple graph representation or expression tree). In contrast, GEP has the characteristics of simple chromosome, linearity and compactness, and easy genetic manipulation. GEP algorithm borrows the concepts of genes, chromosomes, etc. in life science, and uses genetic thought to finish data mining, function discovery, etc. through iteration to obtain the optimal solution of the problem. In genetic evolution, new individuals are obtained through selection, crossover, mutation and evaluation of individual fitness. According to the size of fitness value, the next generation is selected and continuously evolved to finally converge to the optimal solution[20].

In GEP, there are two main bodies: chromosome and expression tree. Genetic information is in chromosomes, while expression tree is the expression of chromosomes. The decoding process from chromosome to expression tree is called translation. GEP gene includes two parts: head and tail. It is encoded by linear symbol strings. These symbol strings are the genotype of the gene. Genes contain multiple gene element bits (coding bits). For a gene, its head length (h) is often defined according to the complexity of the problem, while its tail length (t) and h maintain the following relationship:

$$t = h*(n-1)+1$$

Where n is the largest operation of all functions. Obviously, the length of the gene can be calculated from equation

$$l = h + t = n*h+1$$

The main steps of GEP algorithm are shown in Fig 1[20]. The fitness function is mainly used to calculate the individual fitness value, which is the basis for selection in the evolution process and the "source power" of population evolution. Therefore, the design method of fitness function is the key to the success of evolutionary algorithm in solving the problem. The basic GEP mainly has three fitness functionsOne is that the error type is set to absolute error, and the fitness f_i of individual can be expressed by (3), the other is that the error type is set to absolute error, and the fitness of individual i can be expressed by (4)[21].Among them, M

$$f_i = \sum_{j=1}^{m}\left(M - \left|y_{(i,j)} - y_i\right|\right)$$

$$f_i = \sum_{j=1}^{m}\left(M - \left|\frac{y_{(i,j)} - y_j}{y_j} * 100\right|\right)$$

sets of data come from the sample set, yj is the sample value corresponding to the j-th independent variable,y (i,j) is the return value of chromosome individual i to the fitness sample independent variable j, M is the selection range, the absolute value parts of the two equations correspond to their respective errors, this term is called precision p, therefore, for a perfect adaptation case, $y_{(i,j)}=y_j$, and $f_i = f_{\max} = M*m$ .Since absolute error cannot completely evaluate the accuracy of approximation, and relative error can not only represent absolute error, but also reflect the error characteristics better than absolute error when estimating the error of approximation operation result, the second formula is generally adopted in GEP based on error fitness calculation. For more complex problems, the following fitness function can be used:

$$n \geq \tfrac{1}{2}C_t \text{ then } f_i = n \text{; else}$$

Where n is the number of fitness instances correctly evaluated, C_t is the total number of fitness entities.

GEP algorithm has been well applied in APS(Automatic Problem Solver) software, and a nonlinear model of Hsp90 inhibitor with good prediction ability has been constructed

2.5. BP Neural Network Builds Nonlinear Model

BP-ANN is a type of supervised network, and has been generally accepted for its excellent capability to train rapidly on sparse data sets[22]. BP neural network includes input layer, hidden layer and output layer. The input signal is processed by the hidden layer and then transferred to the output layer. If the output layer cannot obtain the desired output, the error is transferred to reverse propagation, and the neural network weight and threshold are adjusted according to the prediction error to continuously approach the desired output[22]. Sigmoid function and linear function are commonly used in transfer functions of hidden layer and output

2020@ The American Computational Science Society. All rights reserved.

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

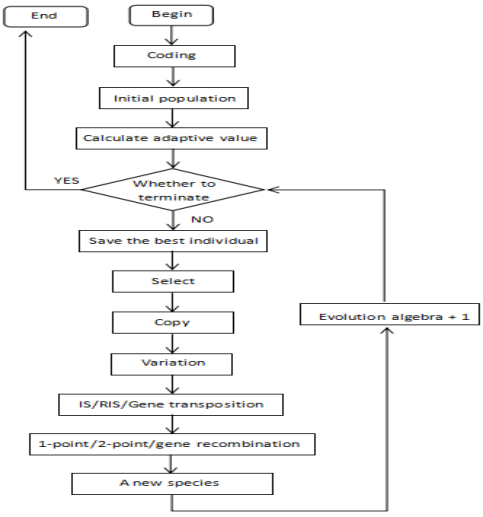


Fig 1. GEP Flow Chart

layer of BP neural network. The Sigmoid function is as follows:

$$F(x)=\frac{1}{1+\exp(-x)}$$

The common three-layer BP artificial neural network (BP-ANN) model is shown in Fig 2, where P_1, P_2... P_n refers to the input parameters, y_1, y_2 ... y_t is the output parameters, w_{ij} is the weight of the neural network, and b is the threshold of the neural network.

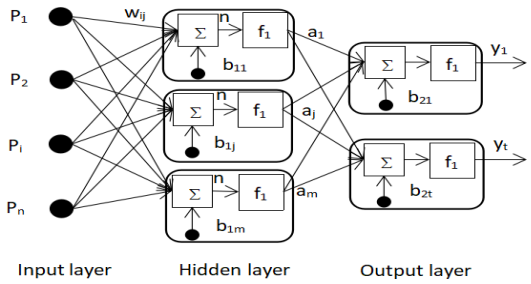


Fig 2.The schematic structure of a BP neural network

The self-learning of BP neural network consists of four processes: the"mode forward propagation" process from the input layer through the hidden layer to the output layer; The"error back propagation" process of correcting the connection weight level by level from the output layer to the input layer for the error signal of the difference between the desired output and the actual output of the network; The convergence of the network is the"learning convergence" process in which the global error of the network tends to a minimum[19]. If the network converges successfully, the required neural network model can be obtained[23].

The implementation steps of the algorithm are as follows[24]:

(1) Initialization weight, learning rate and error target.

(2) Set a smaller non-zero random number for the weights of each layer, and set the learning rate and error target.

(3) input one sample x and corresponding expected output y.

(4) Calculate the actual output and error of each layer according to the formula.

(5) If the error is less than the error target, skip to step (7) and the training is finished.

(6) Calculate the learning error of each layer.

(7) Revise the weight value according to the formula, skip to (3), and recalculate the actual output and error.

(8) Training is over.

3. Results and discussion

3.1. Calculation results of HM

The pEC50 value corresponding to each compound was added to the MOPAC output file, and the file was imported into CODESSA software to obtain 526 descriptors. After screening these molecular descriptors, 7 descriptors were obtained, which could represent the structure and physical and chemical properties of molecules relatively comprehensively. The linear model of 1-9 descriptors established by HM was shown in Fig 3. We could see that after the number of descriptors increased to 8, the increase rate of R^2 and the decrease rate of S^2 decreased slightly, and because the model with too many descriptors had poor generalization ability, 7 descriptors were finally selected as the best linear model. Descriptors were used to build the linear model and their detailed information were presented in Table 2. At the same time, Table 3 showed the correlation matrix

$$pEC_{50} = 7.921 - 0.021175(WNSA-1) - 173.02\,PMIA - 0.02459\,CIC$$
$$-182.03(HACA-1/TMSA) + 0.2991\,HA - 0.0089819(DPSA-1) - 0.0079544(PPSA-1)$$

of these seven descriptors. The calculation formula was shown in (7):

In this model, R^2=0.7937, F=42.86, S^2=0.1050, R^2_{CV}=0.4090. The observed and predicted values of pEC50 for 89 compounds plotted by HM method were as shown in Fig 4. We could see that although R^2 and S^2 were satisfactory, but R^2_{CV} was not very good. It indicated that the generalization ability of the established linear model was not strong. Therefore, we established a nonlinear model by GEP to obtain a more accurate model.

3.2. Calculation results of GEP

In order to obtain a more accurate model with stronger generalization ability and to ensure the comparability of HM and GEP results, we imported seven descriptors obtained by HM, i.e. the corresponding pEC50 values, into APS software and used GEP to establish a nonlinear

2020@ The American Computational Science Society. All rights reserved.

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

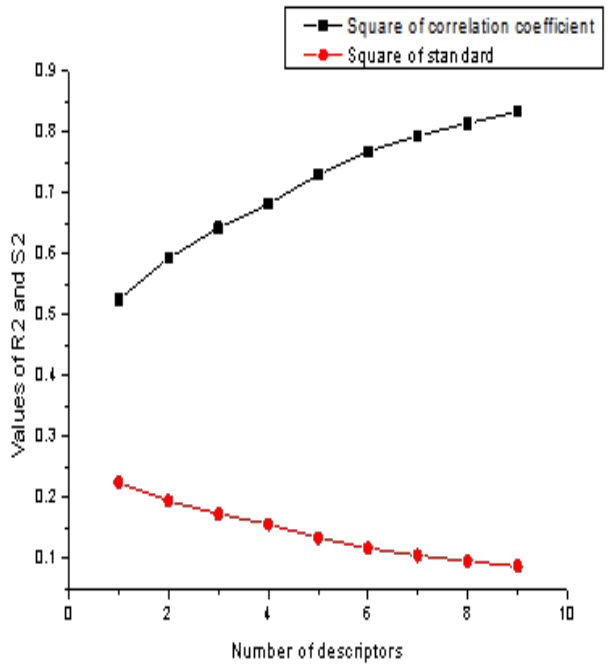


Fig 3. Influences of the number of descriptors on the square of correlation coefficient (R^2) and the square of standard error(S^2) of the regression model

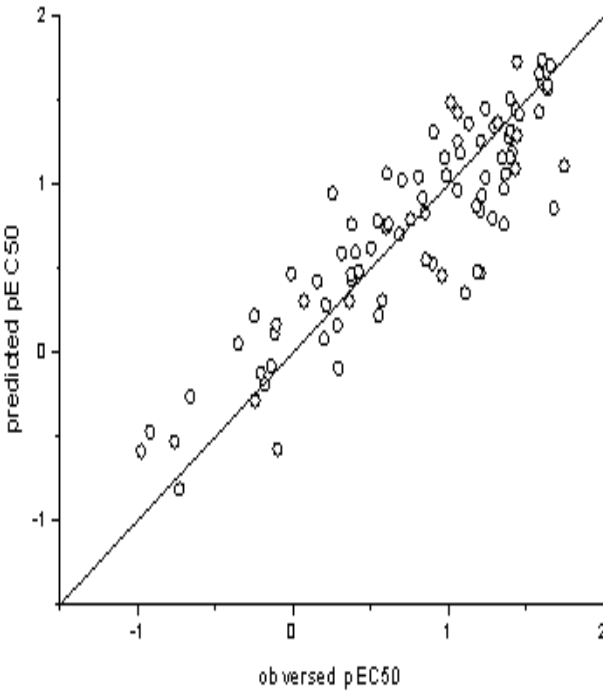


Fig 4. Plot of predicted pEC50 versus experimental values for all 89 compounds by HM

Table 2. The selected descriptors, regression coefficients, DX and t-Test values for the linear model with pEC50 as dependent variable

Descriptors	Chemical means	Confficient	DX	t-Test
Constant	Intercept	7.9210	0.8374	9.5120
WNSA-1	Weighted PNSA(PNSA1*TMSA/1000) [Zefirov's PC]	-0.0212	0.0020	-10.6452
PMIA	Principal Moment of Inertia A	-173.0200	29.0850	-5.9487
CIC	Complementary Information Content (order 2)	-0.0246	0.0040	-6.1638
HACA-1/TMSA	HACA-1/TMSA [Zefirov's PC]	-182.0300	24.7060	-7.3680
HA	HA dependent HDCA-2 [Quantum-Chemical PC]	0.2991	0.0554	5.4001
DPSA-1	Difference in CPSAs (PPSA1-PNSA1) [Zefirov's PC]	-0.0090	0.0017	-5.3653
PPSA-1	Partial positive surface area [Quantum-Chemical PC]	-0.0080	0.0018	4.5205

Table 3. Correlation matrix of the descriptors in the linear model with pEC50 as dependent variable

Descriptors	CIC	WNSA-1	PMIA	HACA-1/TMSA	HA	DPSA-1	PPSA-1
CIC	1	0.3941	-0.6856	-0.7355	-0.5016	0.2755	0.7221
WNSA-1		1	-0.2932	-0.4627	-0.1428	-0.4652	0.3941
PMIA			1	0.6037	0.2301	-0.3637	-0.6856
HACA-1/TMSA				1	0.4948	-0.2641	-0.7355
HA					1	0.1084	-0.5016
DPSA-1						1	0.2755
PPSA-1							1

model. We randomly selected 18 compounds from 89 compounds as test sets and the remained 71 compounds

2020@ The American Computational Science Society. All rights reserved.

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

as train sets. Some parameters in GEP algorithm were adjusted, and a satisfactory result was obtained after evolution: R^2_{train}=0.77, R^2_{test}=0.80. The mathematical expression corresponded to the nonlinear model was as follows:

where x1, x2, x3, x4, x5,x6,x7 represented the

$$pEC_{50} = \sin(\sin(((x_1 + x_1)\%x_4)/\cos(x_3))) + (x_2 * ((1/\cos(x_7 - (x_5 * x_4))) - x_3))$$
$$+\sin(\sin(\sin(\sin((\sin(x_4)) * \lfloor x_1 \rfloor))))) + \cos(((\cos(x_5)\%x_4) * ((1/\cos(x_5)) + (x_6\%x_3))))$$
$$+(\sin((1/\cos((x_5\%x_7))))/(x_5 + (\cos(x_7) - x_4)))$$

descriptors of WNSA-1, PMIA, CIC, HACA-1/TMSA, HA, DPSA-1, PPSA-1 respectively. The function sets used in this model were {'+','-','*','/','mod',' pow',' floor',' sin',' cos',' sec'}, and the other definition parameters were shown in Table 4. Fig 5 showed the predicted pEC50 versus experimental values for train set and test set by GEP. It could be seen that the predicted values of the vast majority of pEC50 were similar to the experimental values, in which the mean square error (MSE) of the train set was 0.10 and the MSE of the test set was 0.11. In comparison, the nonlinear model was better than the linear model.

Table 4. Parameters for the simple symbolic regression problem with EC50 as dependent variable

Parameter names	Values
Number of Chromosomes	100
Number of Genes	5
Function set	+','-','*','/','Mod','Pow','Floor','sin','cos','sec'
Gene head size	8
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
RIS transposition	0.1
Gene transposition rate	0.1
Selection range	100
Precision	0.01

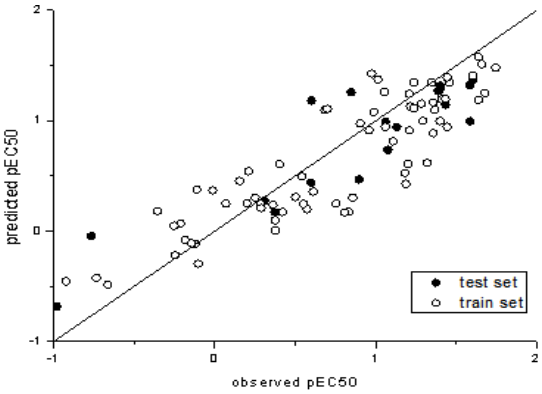


Fig 5. Plot of predicted pEC50 versus experimental values for all 89 compounds by GEP

3.3. Results of BP Neural Network

The GEP model established above had good robustness and prediction ability, but because $R_2$of the model train set was 0.77, slightly less than 0.8, it showed that the model still had certain defects, so the BP neural network was intended to establish a better model. In order to compare with the nonlinear model established by GEP, the established neural network model still used 7 descriptors used by GEP(WNSA-1, PMIA, CIC, HACA-1/TMSA, HA, DPSA-1, PPSA-1) as input layer elements, and the output layer was pEC50 value of aminopyrazole-substituted resorcylate compounds. After many experiments, this study used a 7:4:1 network structure to build the model. In order to have a better comparison with GEP, the same samples as GEP were selected as the test set in BP-ANN method, 10 samples were randomly selected as the verification set in GEP train set, and the remained samples were selected as the train set. In the obtained optimal model, the correlation coefficient of the train set R_{train}=0.9538, the correlation coefficient of the test set R_{tese}=0.9515, and the correlation coefficient of the verification set R_{cv}=0.9551, which were all close to the correlation coefficient (R_{all}=0.9453) of 89 samples. The training function used in the established BP-ANN model was "trainlm". The weight values between the input layer, the hidden layer and the output layer, i.e. the threshold values, were shown in Table 5. The established BP-ANN model was used to predict the 89 samples mentioned above. The observed and predicted values of compound pEC50 were shown in Fig 6. the root mean square error (MSE) of this model was 0.0069. The data of GEP and BP-ANN were compared in Table 6. It could be seen that BP neural network model was obviously better than GEP model and had better fitting effect.

2020@ The American Computational Science Society. All rights reserved.

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

Table 5.Weights and biases of BP-ANN model

	weight							threshold
IW	-1.1672	0.2035	-2.5057	-1.6879	-0.3422	0.3920	0.0333	-0.2023
	0.8756	0.5436	-0.7092	-0.2661	0.1822	-0.1159	-0.9400	0.1509
	0.2396	0.6706	-1.8147	-0.1534	0.6871	-1.4641	-1.3219	0.2730
	-0.3768	0.3879	0.4145	-0.6742	0.7027	-1.1976	-0.1681	-0.7315
LW	0.8931	-1.6709	0.7926	1.0183				

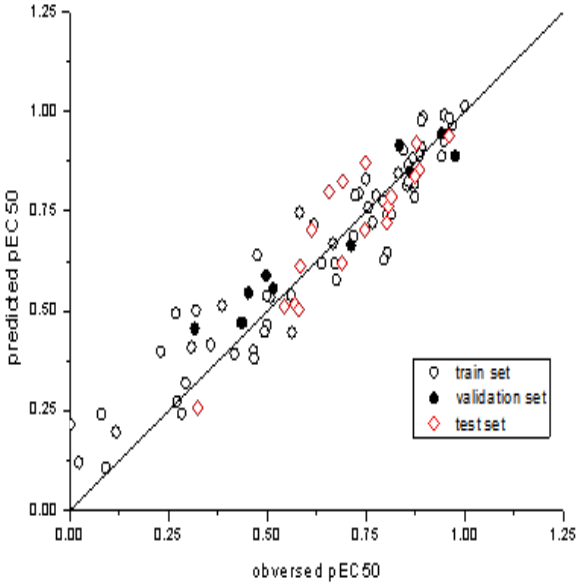


Fig 6.Plot of predicted pEC50 versus experimental values for all 89 compounds by BP-ANN

Table 6.Data comparison between GEP and BP-ANN

Method	Train set		Test set	
	R	MSE	R	MSE
GEP	0.88	0.1000	0.89	0.1100
BP-ANN	0.95	0.0053	0.95	0.0099

4. Conclusion

To sum up, QSAR models for 89 aminopyrazole-substituted resorcylate compounds were established and analyzed. GEP algorithm and BP-ANN algorithm were widely used in two kinds of nonlinear models. However, the model established by artificial neural network has better generalization ability and higher fitting degree than the model established by GEP. BP-ANN is more suitable for solving complicated modeling problems and is a more promising nonlinear modeling method. In addition, the model established above can provide guidance for the design of Hsp90 inhibitors.

References

[1]G.D. Brown, D.W. Denning, N. A. R. Gow. Science Translational Medicine, 4 (2012) 165rv13.
[2]A.J. Alanis. Archives of Medical Research, 36 (2005) 697-705.
[3]C. Walsh, G. Wright. Chemical Reviews, 105 (2005) 391-394.
[4]B. P. P. ChoWdhury, R. Banerjee, R. Paul. Journal of Clinical and Diagnostic Research, 6 (2012) 1577.
[5]C.G. J. McKenzie, U. Koser, L. E. Lewis. Rickettsial Infection and Immunity, 78 (2010) 1650-1658.
[6]L.E. Cowen, S.D. Singh, J.R. Köhler. Proceedings of the National Academy of Sciences of the United States of America, 106 (2009) 2818-2823.
[7]L.E. Cowen, S. Lindquist. Science, 309 (2005) 2185-2189.
[8]Y. Chen, M. Ge, D.J. Chen. Chinese Journal of Antibiotics, 2 (2010) 87-90.
[9]L.E. Cowen. Nature Reviews Microbiology, 6 (2008) 187-198.
[10]X.H. Liu, Z.H. Zhang. Chinese Journal of Pharmaceutical Analysis, 35 (2015) 193-202.
[11]J.C. Chen, L. Qian, Y. Shen. Science in China, Series B: Chemistry, 51 (2008) 111-119.
[12]O.A. Santos-Filho, A.J. Hopfinger. Journal of Computer-Aided Molecular Design, 15 (2001) 1-12.
[13]A. Nanbo, T. Nanbo. QSAR & Combinatorial Science, 21 (2002) 613-616.
[14]X. Wang, Z. Lin, D. Yin. Science in China, Series B: Chemistry, 48 (2005) 246-252.
[15]W.W. Hsieh. Cambridge university press, (2009).
[16]G.Q. Wang, Y. He, D.B. Ren. Computer Study, 3 (2010) 80-83.
[17]D.S. Huang, E.V. LeBlanc, T. Shekhar-Guturja. Journal of Medicinal Chemistry, 63 (2019) 2139–2180.
[18]W. Ren, D.X. Kong. Computers and Applied Chemistry, 26 (2009) 1455-1458.
[19]W.P. Ma. Lanzhou University, (2007).
[20]Y. Zheng. Guangzhou University, (2018).
[21]K.J. Zhang. Zhejiang University, (2010).
[22]X.F. Li, L. Zhao. Chemistry, 78 (2015) 208-214.
[23]M.M. Liu, Y.P. Li, H.Y. Yan. Science and Technology of Food Industry, 20 (2016) 32.
[24]J.L. McClelland, D.E. Rumelhart, The MIT Press, (1986).

2020@ The American Computational Science Society. All rights reserved.