Journal of Computational Science & Engineering 19 (2015) 618-624

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

QSAR study on the IC_{50} of 6-alkenylamides of 4-anilinothieno [2, 3-d] pyrimidine as epidermal growth factor receptor inhibitors in lung cancer

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

[1]Department of Public Health, Qingdao University Medical College, 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, Ningxia Road 308, Qingdao, Shandong 266071, China

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

Abstract: **In this study, a nonlinear quantitative structure–activity relationship model for the prediction of the IC_{50} of 6-alkenylamides of 4-anilinothieno [2, 3-d] pyrimidine as epidermal growth factor receptor (EGFR) inhibitors was developed by the Gene Expression Programming (GEP). This model is based on five descriptors which were selected from all the descriptors by heuristic method (HM). GEP rendered a good correlation between the experimental and predicted logIC_{50} values with a correlation coefficient and a mean error of 0.87 and 0.14 for the training set, 0.71 and 0.41 for the test set, respectively. The results indicate that this QSAR model has good predictive capability and it shows that there is further study we can operate on EGFR inhibitors.**

Keywords: QSAR; Gene Expression Programming; Epidermal growth factor receptor inhibitors; 6-alkenylamides of 4-anilinothieno [2, 3-d] pyrimidine

Received 20 May 2015, Revised 23 July 2015, Accepted 25 July 2015

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

1. Introduction

The mortality of lung cancer is the highest in all cancers and the second cancer that men and women are diagnosed with, which is second only to prostate and breast cancer, respectively [1-3]. Lung cancer is too invasive to be cured; its treatment remains one of challenging tasks in medical industry. Discovery and development of small molecule inhibitors has a more significant impact on the treatment of non-small cell lung cancer than conventional treatments include surgery, radiotherapy and chemotherapy [4]. In the past decade, there are four molecular targeted drugs have been approved for the treatment of lung cancer. They are Gefitinib, Erlotinib, Bevacizumab and Crizotinib [5].

Human Genome Project gave birth to individualized medicine of cancer treatment. Individualized targeted therapy depends on the deep study of the therapeutic target. For the purposes of lung cancer, the increasingly improved epidermal growth factor receptor (EGFR) is one of the most important therapeutic targets [6]. Mediated signal transduction is closely related to the proliferation, existence and adjustment of differentiation of cells. In the surface of cell receptor tyrosine kinase (RTK), EGFR belongs to ErbB family. There are 4 members in EGFR family, EGFR (or ErbB-1), HER-2 (or ErbB-2), HER-3 (or ErbB-3) and HER-4 (or ErbB4) [7]. It has been shown that RTKs plays an important role in tumorigenesis by controlling signal transduction pathways of regulating the proliferation and apoptosis [8]. And EGFR over-expressed in 50% -80% of non-small cell lung cancer patients [9].

With the development and application of computer technology, the studies of QSAR are becoming more and more sophisticated. The range of applications is rapidly expanding involving biological, pharmaceutical science, chemistry environmental science and many other subjects. People hope to understand the relationship between microstructure of compounds from the molecular level and the mechanism between drugs and body so that to design, screen or predict active compounds that people expect. Then infer the factors affecting the present nature and make recommendations for further structural modification. This relies on the successful operation model. In this study, we calculated the 32 molecular structures by the software CODESSA to get five descriptors, which were selected to build several multivariable linear models by the HM. We also use GEP method to analyze and describe the relationship of compound between Enzymatic and cellular inhibitory activities (IC_{50}) to EGFR and molecular structure, and build a stable nonlinear QSAR model. Study shows that compare with the method of HM, the model established by GEP can predict the activity of the compound more excellently and successfully, and developing a new method to the study of the activity of the EGFR inhibitors.

2. Theories and methods

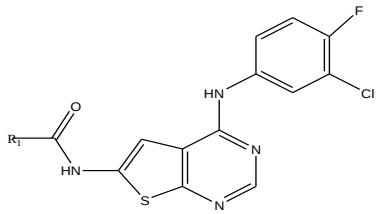
2.1. Data set

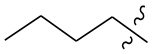
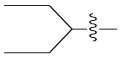
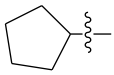
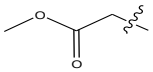
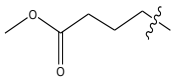
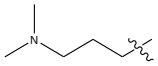
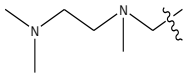
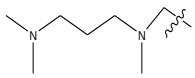
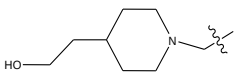
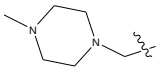
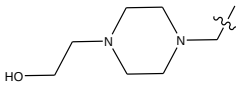
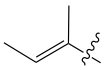
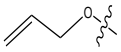
A novel series of 6-alkenylamides of 4-anilinothieno [2,3-d] pyrimidine derivatives were evaluated in terms

2015 © The American Computational Science Society. All rights reserved.

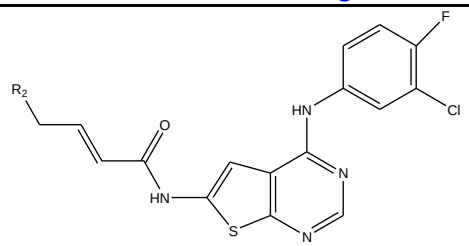
of the anti-proliferative effects on human epithelial carcinoma A431 cells, the biological activities are expressed by the half-maximal inhibitory concentration (IC_{50}) values, and most of the compounds exhibited good potency against EGFR [10]. 41 compounds are divided into 32 compounds of a training set and 9 compounds of a test set randomly, the IC_{50} of the structures are listed in table 1. We build QSAR model according to training set, and then use this model to predict the value of test set.

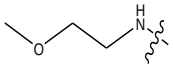
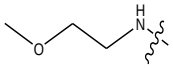
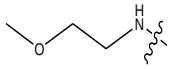
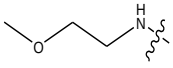
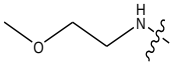
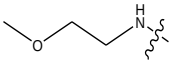
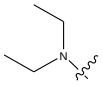
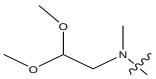
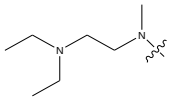
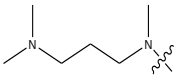
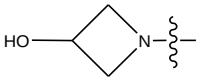
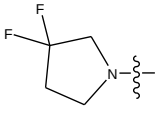
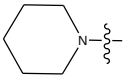
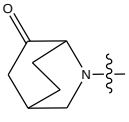
Table 1 The structures, experimental and predictive IC_{50} values of the compounds (GEP)



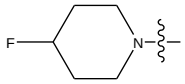
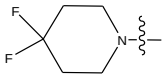
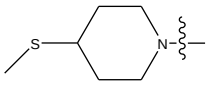
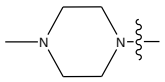
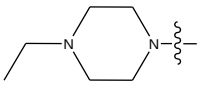
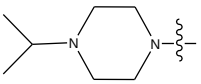
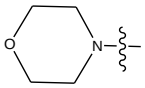
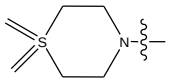
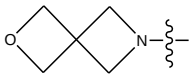
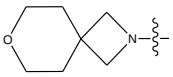
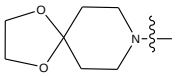
Compounds	R_1	IC50(exp.)	IC50(pre.)	Resid.
8a*		14.21	22.05	7.84
8b		22.82	19.79	3.03
8c		14.94	18.09	3.15
8d		21.33	23.01	1.68
8e		27.84	31.48	3.64
8f*		3.84	1.57	1.91
8g		1.59	2.98	1.39
8h		2.38	2.56	0.18
8i		9.48	2.56	6.92
8j		1.70	4.03	2.33
8k*		10.23	6.02	4.21
8l		0.73	0.25	0.98
8m		24.18	22.80	1.38
8n		26.58	24.34	2.24
8o		3.58	3.93	0.35

2015 © The American Computational Science Society. All rights reserved.



Compounds	R_2	IC_{50}(exp.)	IC_{50}(pre.)	Resid.
9a*		4.85	1.60	3.25
9b		25.4	33.84	8.44
9c		19.19	13.75	5.44
9d		3.93	6.05	2.12
9e		2.21	4.94	2.73
9f*		3.24	4.32	1.08
9g		0.25	1.65	1.40
9h		0.64	1.02	0.38
9i		0.43	1.73	1.30
9j		4.22	1.30	2.92
9k*		3.41	1.82	1.59
9l		1.65	2.12	0.47
9m		3.69	4.06	0.37
9n		0.13	2.33	2.20
9o		1.54	1.62	0.08

2015 © The American Computational Science Society. All rights reserved.

9p*		2.4	3.53	1.13
9q		1.2	1.54	0.34
9r		1.13	1.97	0.84
9s		4.47	2.41	2.06
9t		4.37	1.61	2.76
9u*		4.78	0.44	4.34
9v		2.17	1.81	0.36
9w		8.72	2.86	5.86
9x		0.86	4.02	3.16
9y		13.37	4.31	9.06
9z*		0.55	1.94	1.39

* test set

2.2. Calculation of descriptors

Compounds are often represented by the molecular descriptors in order to obtain a QSAR model. The most important factor affecting the quality of the QSAR model is the numerical representation (often called molecular descriptor) of the chemical structure. The calculation process of the molecular descriptors is described as below:

All of the molecules are drawn into Hyperchem and preoptimized using the MM+ molecular mechanics force field. Then a more precise optimization is done with the semiempirical AM1 method in MOPAC. Calculating until the root mean square gradient reaches 0.01.

The MOPAC output files are introduced to CODESSA program to calculate the descriptors; Totally there are eight groups of 593 descriptors in CODESSA: Constitutional, Topological, Geometrical, Electrostatic, Quantum-chemical, Thermodynamic [13-14].

2.3. Development of linear model by HM

Once the molecular descriptors are generated, the HM in CODESSA is used to pre-select the descriptors and build a linear model.

The advantages of HM are: high speed and no software restrictions on the size of the data set. Also, it can estimate the quality of correlation from the known data quickly and obtain some vintage regression models. The details of selecting descriptors can refer to the literature. The goodness of the correlation is tested by the square of coefficient R egression (R^2), square of cross-validate coefficient regression (R^2_{CV}), the F-test (F), and the standard deviation (S^2). Finally, five descriptors are selected from descriptors pool and the linear model is produced by the HM.

However, the factors that influence the IC_{50} of compounds are complex, and more, not all of them are in linear correlation with the IC_{50}. So the non-linear methods by GEP are used to develop more accurate models.

2.4. Development of nonlinear model by the GEP algorithm

Gene Expression Programming was invented by Ferreira in 1999 [11], and it is the natural development of genetic algorithms and the genetic programming (GP). The GEP uses the same kind of diagram

2015 © The American Computational Science Society. All rights reserved.

representation of the GP, but is expressed with genes, which is the difference in between. It uses character linear chromosomes composed of genes structurally organized in a head and a tail. The GEP includes chromosomes and the expression trees (ETs). The processes of gene encoding and translation are simple. And the relationship between the symbols of the chromosome and the functions and terminals they represent is corresponding.

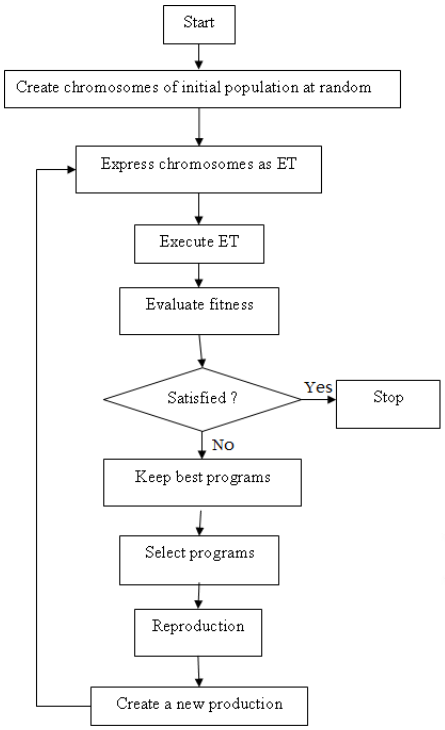


Figure 1. The brief flow chart of GEP.

The process of the GEP has five steps, which are shown in Figure 1. Use automatic problem solver (APS) [12] program to get the model result.

3. Results and discussion

3.1. The HM model

593 descriptors were calculated by the CODESSA program totally for all the compounds. HM can not only filter a group of molecular descriptors, but also build a linear model according to the descriptors. Use R^2, R^2_{CV}, F, and the S^2 to test quality of the correlation. And R^2 and R^2_{CV} improved, S^2 decreased with the growing of number of molecular descriptors (Figure 2).

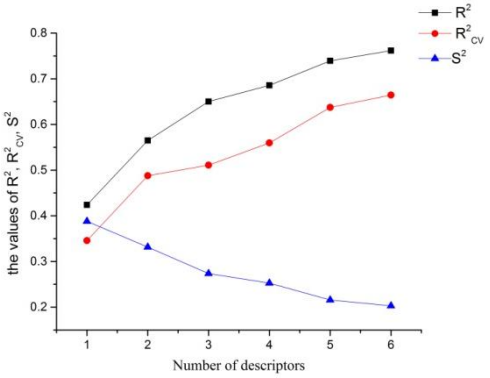


Figure 2. Influence of the number of descriptors on the correlation coefficient, the cross-validation correlation coefficient and the square error.

It should be noted that, excessive molecular descriptors may cause over-fitting [13]. The details of five descriptors are given in Table 2. After a comprehensive considerationthe, the quality of the correlation is best when the number is five. Taking R^2_{CV} into account, the 5-parameter model should be considered as the best linear model (Figure 3.) and the corresponding descriptors were applied as inputs for the non-linear mode.

Table 2 The grouping of five descriptors

Symbol	NAME	GROUP
MERIN	Min electroph React index for a N atom	Electrostatic
MPCO	Max partial charge for a O atom [Zefirov's PC]	Electrostatic
KSI	Kier shape index (order 3)	Topological
HDSA	HA dependent HDSA-2 [Quantum-Chemical PC]	Quantum-chemical
TM2RE	Tot molecular 2-center resonance energy of atoms-Quantum	Quantum-chemical

To avoid the possible multi-collinearity of molecular descriptor scales, Table 3 lists the correlation coefficients between the descriptors were selected. From the table we can see that all the descriptors are strongly orthogonal. The model has a good statistical reliability.

$$IC_{50} = 21589MERIN + 0.91566KSI + 0.99533HDSA + 712.84MPCO + 7.0521TM2RE \quad (1)$$

$R^2=0.74$ $R^2_{CV}=0.64$ $S^2=0.216$ $F=19.83$

Min electroph React index for a N atom (MERIN) and Max partial charge for a O atom [Zefirov's PC] (MPCO): These descriptors reflect characteristics of the charge distribution of the molecule. The empirical partial charges in the molecule are calculated using the approach proposed by Zefirov [87MI883, 87ZOK4]. This method is based on the Sanderson electronegativity scale and uses the concept which represents the molecular electronegativity as a geometric mean of atomic electronegativities.

2015 © The American Computational Science Society. All rights reserved.

Kier shape index (order 3) (KSI): The shape of molecule depends on the number of skeletal atoms, the molecular branching and the special parameter which is calculated as the ratio of the atomic radius (ri) and the radius of the carbon atom in the sp3 hybridization state (ro) [90MI151].

HA dependent HDSA-2 [Quantum-Chemical PC](HDSA) and Tot molecular 2-center resonance energy of atoms-Quantum (TM2RE): Numerous studies have involved descriptors calculated from quantum chemical data [79MI467, 81MI157, 83AC738, 84MI115, 87MI335, 89JA3783, 90QSAR107, 91MI247]. The quantum-chemical descriptors add important information to the conventional descriptors. A large varieties of quantum-chemically calculated descriptors are available in the CODESSA program. These quantum-chemical descriptors are calculated proceeding from the output results of the MOPAC program package [89MI1], provided that the necessary keywords were used in the quantum calculation.

All the five descriptors are positively correlated with IC_{50}.

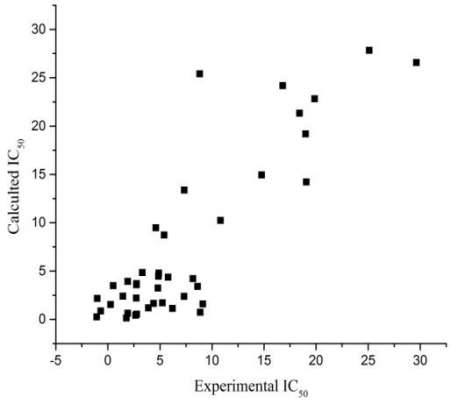


Figure 3. Plot of experimental and calculated IC50 by HM.

Table 3 Correlation matrix of the descriptors in the model

	MERIN	KSI	HDSA	MPCO	TM2RE
MERIN	1.000	-0.716	-0.681	0.378	-0.619
KSI		1.000	0.489	-0.072	0.737
HDSA			1.000	-0.211	0.294
MPCO				1.000	-0.405
TM2RE					1.000

3.2. The GEP model

The five molecular descriptors and activity data (IC_{50}) are introduced to build a nonlinear model by APS software. APS is used to predict IC_{50} because it allows an easy optimization of intermediate solutions and easy test of the evolved models against a test set. The simplest function set is expressed as F={+, -, *, /, Abs, Sec, Asech }.

A small population of 46 individuals is calculated by [14]

$$f_{fitness}(i) = 1000 \times 1/(E_i + 1) \quad (2)$$

and

$$E_i = \frac{1}{n}\sum (P_{(ij)} - T_j)^2 \quad (3)$$

In the run, the 88th generation is a very good solution, with a R^2 of 0.87 for the training set and a R^2 of 0.71 for the test set. The ETs can be converted into Eq. (4).

$$y = fabs(x_2 - \frac{x_1 x_2 x_3^2}{x_4}) + \frac{1}{\cos(\frac{x_3 - x_4 x_5}{x_4^2 x_5})x_5} + \frac{x_3 + fabs(\frac{2}{\cos x_3})}{x_2} + [x_5 - x_1(x_3 + x_5)]x_3(x_3 + x_5) + \frac{1}{\cos[\frac{1}{\cos(x_3 - x_4)x_3}]x_2} + \frac{x_3}{x_2} \quad (4)$$

In Eq.(4), x_1 , x_2 , x_3 , x_4 and x_5 represent variables of MERIN, KSI, HDSA, MPCO and TM2RE, respectively. The fitting situations of the training and the test set for the GEP model are shown in Figure 4 and 5 respectively. Most of the predicted values are very close to the experimental ones.

2015 © The American Computational Science Society. All rights reserved.

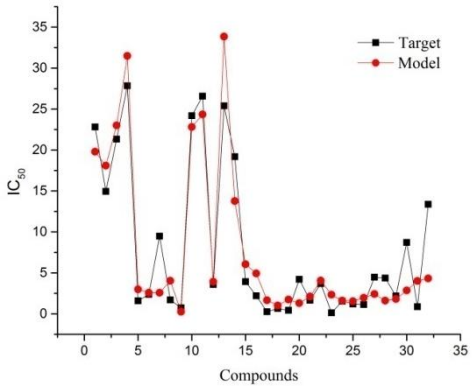


Figure 4. Fitting curve of experimental and calculated IC50 in training set by GEP.

3.3. The GEP compares with HM

In the GEP method, the R^2 and S^2 are 0.87 and 0.14 for the training set, 0.71 and 0.41 for the test set, respectively. And the R^2 and S^2 are 0.87 and 0.35 for all compounds. From the results we can see that the GEP method is better than the HM method.

The above discussion shows that the physical chemistry property of compounds is an important factor in influencing the value of IC_{50}. The selected five descriptors are the most important physical chemistry properties for either the construction of QSAR model or the prediction of IC_{50} of 6-alkenylamides of 4-anilinothieno [2, 3-d] pyrimidine.

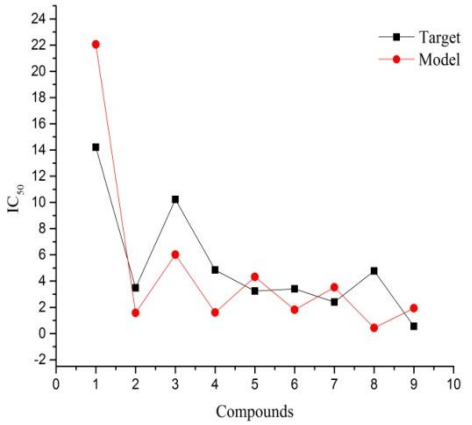


Figure 5. Fitting curve of experimental and calculated IC50 in testing set by GEP.

4. Conclusion

In this study, we build the model By GEP method and use it to predict the biological activities of these compounds significantly, suggesting that 6-alkenylamides of 4-anilinothieno [2, 3-d] pyrimidine as epidermal EGFR inhibitors is worthy of study to treat lung cancer. Although there are deficiencies when use GEP method. For example, there should be enough and accurate original data set to get a more optimal model. The results of this study shows that GEP is an effective and accurate tool for building a nonlinear model, and GEP method has a better prediction ability than that of HM. It also shows that we could measure properties of part of a series substance by experiment and then build QSAR model to obtain physical and chemical properties of substances, so that we could predict the properties of others, which saves time and reduces the cost. In the aspects of predicting biological activity of the new synthetic drugs, using these methods to build QSAR model, predict the efficacy of new drugs, and provide references to filter and synthesize drugs.

References

[1] R.S. Herbst, J.V. Heymach, S.M. Lippman, Lung cancer. N. Engl. J. Med., 359 (2008) 1367-1380.
[2] http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-029771.pdf.
[3] Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, CA. Cancer J. Clin., 61 (2011) 69-90.
[4] L.V. Sequist, T.J. Lynch, Annu. Rev. Med., 59 (2008) 429-442.
[5] P.T. Cagle, L.R. Chirieac, Arch. Pathol. Lab. Med., 136 (2016) 504-509.
[6] K. Wu, L. House, W. Liu, International journal of molecular sciences, 13 (2012) 11471-11496.
[7] T. Mitsudomi, Y. Yatabe, Cancer Sci., 98 (2007)1817-1824.
[8] S.V. Sharma, D.W. Bell, J. Settleman, D.A. Haber, Nat. Rev. Cancer, 7 (2007)169–181.
[9] E.N. Martine, M.S. Hahn, G.W. McKenna, In Advances in Radiation Oncology in Lung Cancer, 5 (2014) 66-71.
[10] X. Ji, T. Peng, X. Zhang, Bioorganic & medicinal chemistry, 22 (2014) 2366-2378.
[11] C. Ferreira, J. Com. Sys., 13 (2001) 87-129.
[12] http://www.gepsoft.com/gepsoft.
[13] J.Z. Li, Lanzhou University, 2009.
[14] http://www.gene-expression-programming.com.

2015 © The American Computational Science Society. All rights reserved.