Journal of Computational Science &Engineering 44 (2019) 1094-1102

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

Quantitative structure activity relationship study of hexacyclic skeletons compounds based on gene expression programming

Kai Leng[1], Peijian Zhang[1], Hongzong Si[2], Honglin Zhai[3]

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

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

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

Abstract: **The quantitative structure-activity relationship models of 33 tetrahydroisoquinoline–tetrahydrocarboline hybrid skeletons derivatives were established with gene expression programming (GEP) extended from genetic algorithms and genetic programming. The QSAR model predicts antitumor activity through descriptors. These descriptors were calculated in CODESSA software and selected from descriptors' pool based on heuristic method. And 4 descriptors were selected to establish linear regression (MLR) model. A comparison between the result of proposed linear and nonlinear regression showed the superiority of QSAR modeling by GEP method over MLR.**

Keywords: Antitumor activity; Heuristic method; Quantitative structure activity relationship; Heuristic method; Gene expression programming

Received 3 October 2019, Revised 7 December 2019, Accepted 10 December 2019

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

1. Introduction

Bistetrahydroisoquinoline alkaloids are a large family of natural products possessing complex structures and exhibiting potent biological activities [1]. The tetrahydro-β-carboline (THβC) ring system has potent antitumor activity and biosynthetic pathway similar to that of tetrahydroisoquinoline (THIQ) (the Pictet–Spengler reaction) [2-6]. Studying the hexacyclic skeletons compounds with QSAR can help us further explore its antitumor activity, and would be helpful to further understand the structure–activity relationship of these kinds of natural products. Quantitative structure activity relationship (QSAR) can significantly reduce the time and resources required for studying the activity of compounds.

The resultant toxic effects of chemical compounds on the human tumor cells exhibit certain specificity which depends on the compound's chemical structure and reactivity [7]. Molecular descriptors and chemometrics are a powerful combined tool for studying pharmaceutical, toxicological, and environmental problems. Compounds with similar molecular structures should have similar physicochemical properties, environmental fate, and toxicological effects [8]; that is, there are inherent relations between molecular structures and their physicochemical properties, and environmental behavioral and toxicological parameters. Heuristic algorithm can help us find the suitable descriptors.

The QSAR model predicts physical and chemical properties, environmental fate, and toxicological effects by establishing linear or non-linear mathematical models using chemical descriptors [9]. The biological activity of antitumor compounds can be expressed as a half maximal inhibitory concentration (IC50), which is the concentration of an inhibitor required for fifty percent inhibition of cells, enzymes, cell receptors or microorganisms [10].

In order to better predict its physical and chemical properties, the GEP algorithm is proposed. GEP is a novel adaptive evolutionary algorithm overcoming some shortcomings of genetic algorithms (GA) and genetic programming (GP). According to research that GEP is 100-60,000 times faster than traditional GA and GP [11].

2. Experimentation

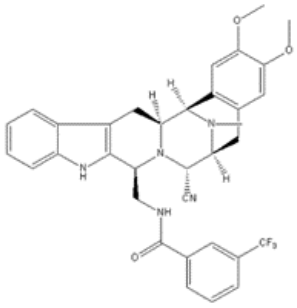
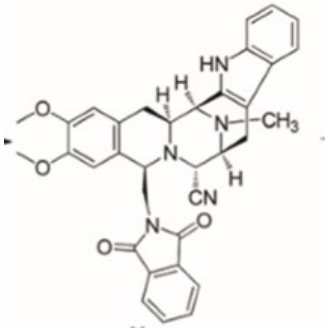
2.1. Datasets and descriptor calculation

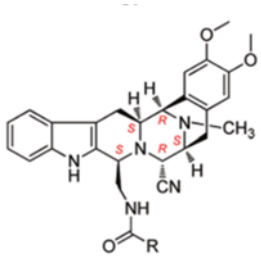
Data tetrahydroisoquinoline-tetrahydrocarboline hybrid skeletons derivatives that were taken from existing literature [12] and given in table 1. And in order to better analyze that, lg(IC50) values were employed as the dependent variables and shown in table 1. In the GEP method step, the dataset was randomly partitioned into a training set of 23 compounds and a test set of 9 compounds. The training set was employed to developed the model, and the test set was employed to assess the prediction ability of model.

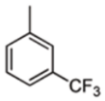
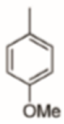
QSAR model uses descriptors to represent biochemical properties of compounds. In this case, the generation and calculation were particularly important. And the calculation steps of the molecular descriptors were described as follows: At first, the molecular descriptors of 32 compounds were drawn in ChemDraw software. The 2D structures headed to energy minima that were accomplished by using AM1 method [13] of HyperChem7.5 [14]. All of the optimized molecules were put into MOPAC 6.0 software. Finally. the results were imported into CODESSA software[15] to calculate five classes of molecular descriptors: constitutional, geometrical, topological, electrostatic and quantum-chemical.

2019@ The American Computational Science Society. All rights reserved.

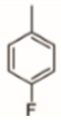
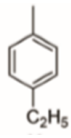
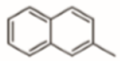
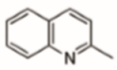
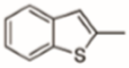
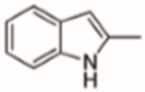
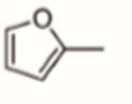
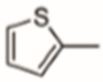
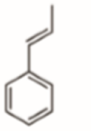
Table 1. Experimental and calculated lg(IC50) by GEP of 32 compounds

No.	Compound	lg(IC50)	Predict	Residual
1		1.8887	1.742	0.1466
2		2.3424	2.3527	0.0101

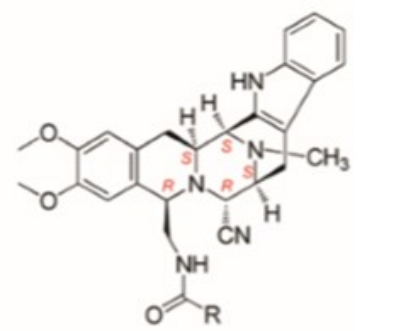


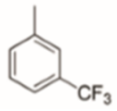
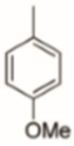
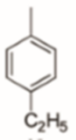
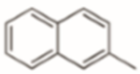
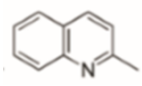
No.	R	lg(IC50)	Predict	Residual
3*		1.9449	1.1528	0.792
4*		1.6532	1.6994	0.0462
5*		2.1702	0.9311	1.239
6		0.1205	0.8498	0.7293

2019@ The American Computational Science Society. All rights reserved.

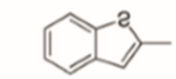
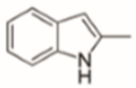
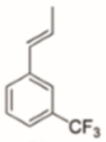
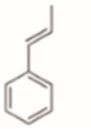
7		**2.3283**	**1.7739**	**0.5543**
8		**1.2787**	**1.2287**	**0.0499**
9		**1.2405**	**0.8461**	**0.3943**
10		**1.1205**	**1.1336**	**0.0131**
11		**2.5514**	**2.4037**	**0.1477**
12		**2.4698**	**2.6329**	**0.1631**
13		**1.7723**	**2.865**	**1.0927**
14		**2.4899**	**2.7701**	**0.2802**
15		**1.9405**	**24,353**	**0.4948**
16		**2.2174**	**2.6643**	**0.4469**
17		**2.534**	**2.3803**	**0.1536**

2019@ The American Computational Science Society. All rights reserved.



No.	R	lg(IC50)	Predict	Residual
18*		3.8536	3.7082	0.1454
19		3.8475	3.4958	0.3517
20		3.6739	3.5935	0.0804
21		3.8692	3.0549	0.8143
22		3.5977	3.6225	0.0248
23		3.2787	3.6349	0.3562
24		3.8573	3.7599	0.0973
25		3.4425	3.6991	0.2456

2019@ The American Computational Science Society. All rights reserved.

26		**3.8469**	**3.0185**	**0.8283**
27		**3.8954**	**3.4736**	**0.4218**
28*		**2.9243**	**2.6386**	**0.2857**
29*		**3.2504**	**3.1576**	**0.0928**
30*		**2.9444**	**2.4743**	**0.4701**
31*		**3.4886**	**3.3662**	**0.1223**
32*		**3.8376**	**2.9608**	**0.8767**

***The compounds of the test set.**

2.2. Establishment of linear model

After all molecules descriptors are calculated using MOPAC software, the stepwise multiple linear regression procedure was used for model generation, and the Heuristic method (HM) [16-18] in CODESSA software was used to select appropriate descriptors. In the stepwise multiple regression, all descriptors are treated as variables. At first, an equation is produced without any variables. Then, each variable is added to the equation, and a new regression was established by HM and sorted by their coefficient regression. At the same time, the square of coefficient regression (R^2) was preserved. Finally, the most suitable linear regression model was selected after comparison, equal to the most appropriate descriptor combination was selected. Furthermore, the square of coefficient regression (R^2) , square of cross-validate coefficient regression (R^2_{CV}), the F-test (F) , and the standard deviation (S^2) was used to test the correlation.

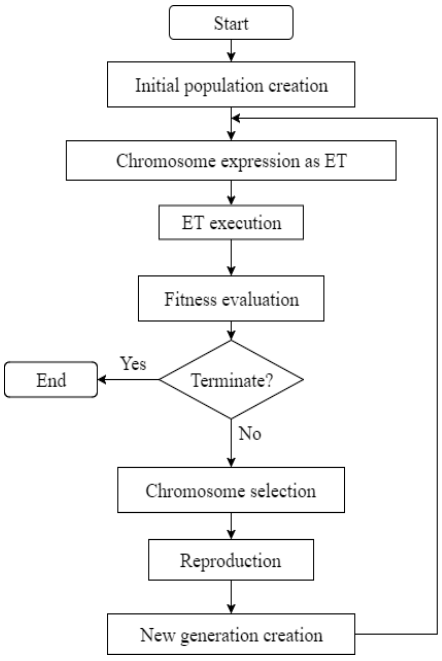


Figure 1. Flowchart of GEP algorithm.

2019@ The American Computational Science Society. All rights reserved.

2.3. Establishment of nonlinear model

GEP algorithm is developed from genetic algorithms (GAs) and genetic programming (GP). While absorbing the advantages of both, overcomes the shortcomings of both. Its distinctive feature is the use of simple coding to solve complex problems. The GEP algorithm is described in detail in this article [19]. And the basic algorithm processes were summarized and shown in Figure 1. The basic GEP algorithm steps are as follows: Step 1. Initialize the population. Step 2. Translating all chromosomes into corresponding expression trees. Step 3. calculate the fitness. And exit if the termination condition is met. Step 4. Selection. Step 5. Select chromosomes based on fitness. Step 6 reproduce the next generation and back to the step 2.

The GEP gene consisted of two parts, the head and the tail, and the gene contained multiple gene element bits. The values of the gene element bits were taken from the set of terminal T and the set of function F [20-21].

3. Result and discussion

3.1. The results of HM

537 descriptors were calculated for each compound by the CODESSA software. To find the most suitable combination of descriptors, multiple linear regression equations with numbers from 1 to 10 were calculated. The prediction ability of the evaluation model is evaluated by R2, Rcv2, and S2, and the result is shown in figure2 and figure 3. Finally, the model with the number of 4 descriptors was found to be the most suitable, as shown in table2 in detail. In order to make R2, Rcv2, and s2 better expressed in a table, logarithmic processing is performed on S2。 The results show that as the number of descriptors increases, R2, Rcv2 gradually increase, and S2 gradually decreases.

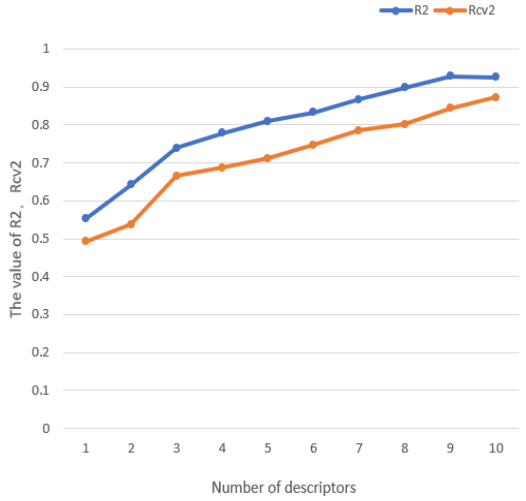


Figure 2. Plot of measured and calculated log(IC_{50}) by HM

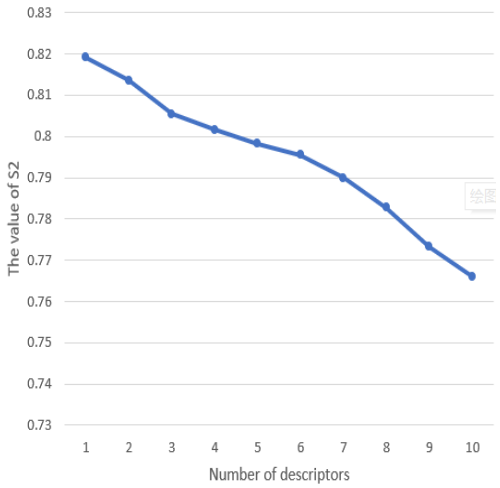


Figure 3. Plot of measured and calculated log(IC_{50}) by HM

Because the regression equation is obtained using stepwise regression, there is generally no multi-collinearity between the obtained molecular descriptors. But for the accuracy of the results, the correlation coefficients between the descriptors were calculated showing in Table 3. The results show that the correlation coefficient of any two molecular descriptors does not exceed 0.8, indicating that all descriptors were independent. Therefore, the constructed linear model has strong statistical reliability. And the plot for the HM model is shown in Figure 4.

The linear model equation is as follows:

$$lg(IC50) = (1.3398e+07) + (-2.3041e+08)*MPCN + (4.5210e+04)*RNA + (-7.7890e+04)*MRIN + (-2.3919e+04)*RNC$$

2019@ The American Computational Science Society. All rights reserved.

Table 2. The selected molecular descriptors, and their physical-chemical meaning, coefficient, and T-test.

Symbol	Physical-chemical Meaning	Coefficient	T-test
MPCN	Max partial charge for a N atom [Zefirov's PC]	-2.3041e+08	-6.75
RNA	Relative number of aromatic bonds	4.5210e+04	4.68
MRIN	Max 1-electron react. index for a N atom	-7.7890e+04	-4.12
RNC	Relative number of C atoms	-2.3919e+04	-4.13

Table 3. Correlation matrix of the 5 descriptors

	MPCN	RNA	MRIN	RNC
MPCN	1.00	-0.02	0.31	0.55
RNA		1.00	0.30	0.09
MRIN			1.00	-0.54
RNC				1.00

Figure 4. Plot of measured and calculated lg(IC_{50}) by HM.

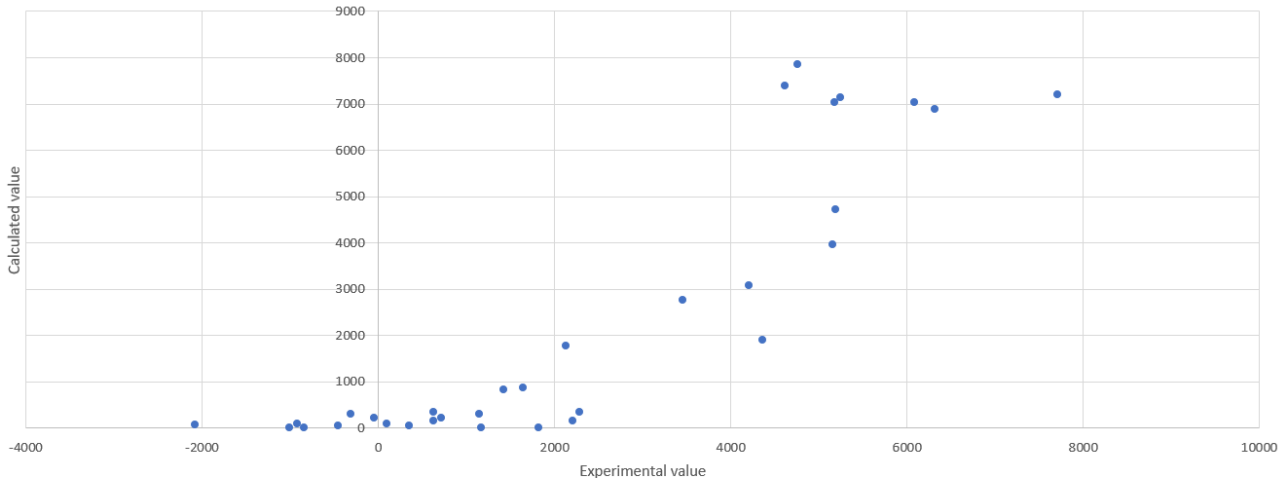


Table 4. Parameters for the simple symbolic regression problem

Parameter Names	values
Number of Chromosomes	100
Head Size	8
Number of Genes	5
Linking Function	+
Generations without Change	200
Number of Tires	3
Max. Complexity	5
Selection Range	100
Precision	0.01
0/1 Rounding Threshold	0.1
Mutation rate	0.044
Inversion rate	0.1
IS Transposition rate	0.1

2019@ The American Computational Science Society. All rights reserved.

RIS Transposition rate	0.1
1-Point Recombination rate	0.3
2-Point Recombination rate	0.3
Gene Recombination rate	0.1
Gene Transposition rate	0.1
Function sets	+, -, *, /, Log, Abs, Acos,Sin

3.2. The results of GEP

The data set is divided into a training set of 23 compounds and a test machine for 9 compounds, and then a software automatic problem solver program (APS) that implements the GEP algorithm is used to construct a nonlinear model. The parameters used in the evolution process were listed in Table 4. The table lists the parameters used in the evolution process. Satisfactory results were obtained in the 328th generation, with correlation coefficients of 0.82 and 0.75, and root mean square error (RMSE) of 0.45 and 0.6, respectively. The predicted results are shown in Figure 6. The fit results of the training and test sets are shown in Figures 6 and 7. Moreover, the nonlinear model equation decoded by the ETs was as follows:

$$\begin{aligned} \lg(IC50) &= fabs(\sin(x_1 - x_2)) \\ &+ \arccos\left(log_{10}\left(fabs(x_4 + fabs(x_1)/fabs(x_4))\right)\right) \\ &+ (log_{10}(x_4 * sinx_3) \\ &+ \sin\left(((sinx_1) - log_{10}(x_2)) + ((x_3 * x_2) - \arccos x_1)\right) \\ &+ fabs(\sin(x_4 - x_1)) \end{aligned}$$

,

where x1, x2, x3 and x4 represented variables of MPCN, RNA, MRIN and RNC respectively.

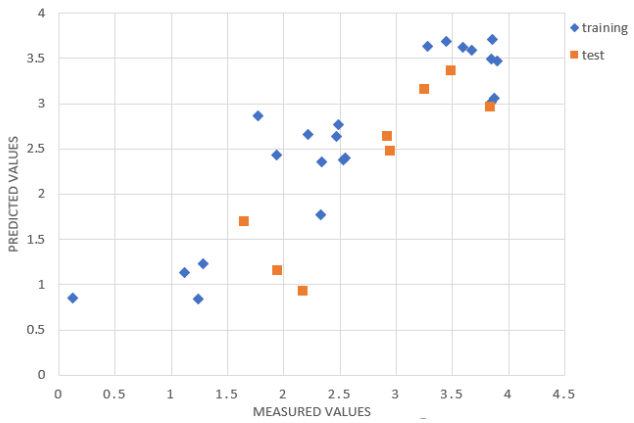


Figure 5. Plot of measured and calculated lg(IC_{50}) by GEP.

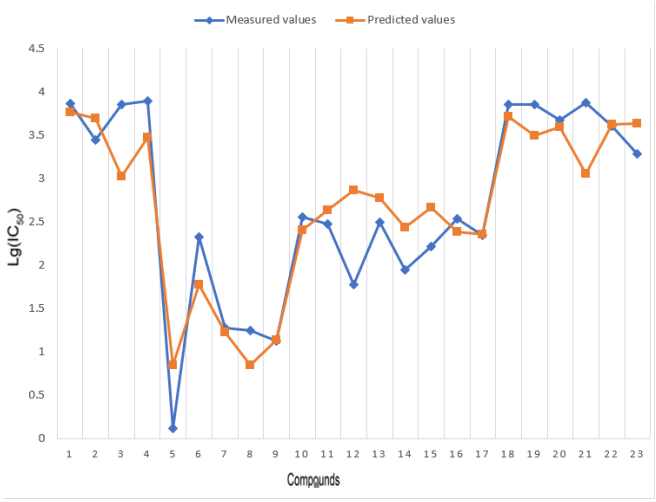


Figure 6. Fitting curve of measured and calculated lg(IC_{50}) in training set by GEP.

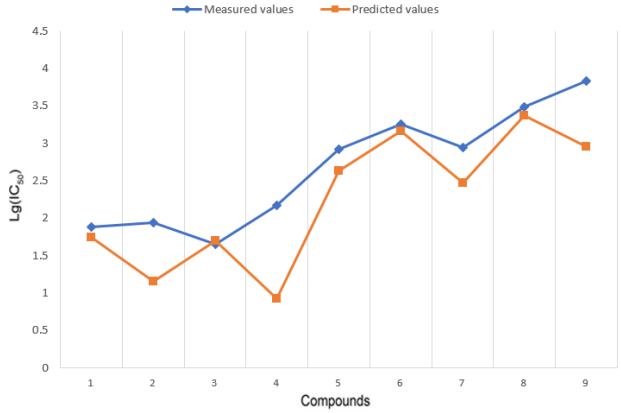


Figure 7. Fitting curve of measured and calculated lg(IC_{50}) in test set by GEP.

4. Conclusion

A quantitative model has been developed to predict the THIQ compounds by GEP in this work. Compared with the linear prediction model established by HM, the nonlinear model established by GEP has better prediction ability and stability. Therefore, the GEP offers a new method for solving more complex technological and scientific problems. This provides guidance for further research on the design and synthesis of THIQ compounds

Conflict of interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

The authors are very grateful to the APS software providers.

References

2019@ The American Computational Science Society. All rights reserved.

[1] J. D. Scott and R. M. Williams, Chem. Rev., 2002, 102, 1669–1730.
[2] E. D. Cox and J. M. Cook, Chem. Rev., 1995, 95, 1797–1842.
[3] J. Stöckigt, A. P. Antonchick, F. R. Wu and H. Waldmann, Angew. Chem., Int. Ed., 2011, 50, 8538–8564.
[4] A. E. Laine, C. Lood and A. M. P. Koskinen, Molecules, 2014, 19, 1544–1567.
[5] J. Jaworska, J. Walker, C. Watts, A. Worth. Environmental Health Perspectives, 111 (2003) 1391-1401.
[6] A.K. Bandyopadhyaya, J. Johnsamuel, A.S. Al-Madhoun, S. Eriksson, W. Tjarks, Bioorg. Med. Chem, 13 (2005) 1681–1689.
[7] M.T. Sacan, S.S. Erdem, G.A. Ozpinar, I.A. Balcioglu. Journal of chemical information and computer sciences, 44 (2004) 985-992.
[8] Nel, Xia, Mädler, &Li, 2006.
[9] C. Ferreira, Adv. Complex Syst, 5 (2002) 389–408.
[10] R.G. Karki, V.M. Kulkarni, Bioorg. Med. Chem, 9 (2001) 3153–3160.
[11] Guven, A. Aytek. Programming. J Hydrol Eng, 14 (2009) 812-832.
[12] Xiangran Lu, Xuan Pan, Baohe Guan and Zhanzhu Liu 10.1039/c9ob02426c.
[13] J. Mancuso, R.J. McEachern. J Mol Graph Model, 15 (1997) 82-90.
[14] M. Froimowitz. BioTechniques, 14 (1993) 1010-1013.
[15] A.R. Katritzky, S. Perumal, R. Petrukhin, E. Kleinpeter. Journal of chemical information and computer sciences, 41 (2001) 569-574.
[16] A.R. Katritzky, V.S. Lobanov, M. Karelson. Chem. Soc. Rev, 24 (1995) 279–287.
[17] H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan. Analytica Chimica Acta, 558 (2006) 86–93.
[18] H.X. Liu, C.X. Xue, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci, 44 (2004) 161–167.
[19] C. Ferreira. Complex Syst, 13 (2001) 87–129.
[20] L. Teodorescu, D. Sherwood. Computer Physics Communications, 178 (2008) 409–419.
[21] W.M. Shi, X.Y. Zhang, Q. Shen. European Journal of Medicinal Chemistry, 45 (2010) 49–54.

2019@ The American Computational Science Society. All rights reserved.