Journal of Computational Science & Engineering 21 (2015) 671-676

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

Quantitative Structure-activity Relationships (QSAR) of a series of 5-HT_{2C} Agonists

Haoyang Sun[1], Ting Zhao[2], Xuejing Zhang[2], Fucheng Song[3]

[1]Department of Integrated Chinese and Western medicine, the Medical College of Qingdao University, Qingdao, Shandong 266071, China

[2]Department of Nephrology, the Affiliated Hospital of Haici, Qingdao, 266023, China

[3]Department of Public Health, the Medical College of Qingdao University, Qingdao, Shandong 266071, China

Abstract: **Gene expression programming (GEP) as a novel machine learning technique was firstly used to predict the EC50 of 5-HT_{2C} agonist. Each kind of compound was represented by a series of calculated structural descriptors. GEP model is based on five descriptors which were selected from the descriptors' pool by heuristic method (HM) to build multivariable linear model. As a remarkable result, the GEP method generated a nonlinear quantitative model with a correlation coefficient and a variance of 0.82 and 0.10 for the training set, 0.78 and 0.18 for the test set, respectively. It indicates that GEP can be applied to study the 5-HT_{2C} agonist, and the predicted results are well consistent with experimental values.**

Keywords: QSAR; 5-HT_{2C}; Gene expression programming; Heuristic method; Traditional Chinese medicine

Received 29 November 2015, Revised 23 December 2015, Accepted 2 January 2016

1. Introduction

The tonifying kidney and regulating qi method from traditional Chinese medicine can enhance the function of the kidney and spleen and make a person in calm [1,2]. This method can play a role of antidepressant through various channels. Depression is a complex illness, with a range of psychological and social factors as well as genetic and neurobiological factors influencing its development [3-5]. The decrease of 5-Hydroxytryptamine2C (5-HT_{2C}) will promote the happening of the depression [6,7]. The recently studies found that the tonifying kidney and regulating qi method could improve the intracorporal 5-HT_{2C} level of people with depression. It provides a mechanism about the treatment of depression. The prescription has many compounds that are the derivatives of 5-HT_{2C} agonist. Studies found the tetrahydropyrazinoisoquinolinones could be one of the-HT_{2C} agonists.

In this study, our purpose was focused on identification of novel structural motifs and high selectivity with potent 5-HT_{2C} agonist. The quantitative structure-activity relationship (QSAR) method is based on the chemical structure information and using quantitative parameters alone or combining with quantitative conventional parameters to construct the mathematical model. QSAR was employed to find patterns between the biomedical activity and a variety of computationally calculated physical–chemical structural properties. Its techniques have been successfully used in the discovery of anticarcinogen for chemotherapy, including breast cancer, colon cancer, and other anticancer compounds [8-11]. Gene expression programming (GEP) is a newfangled nonlinear regression method that will help us get better QSAR model. In this study, GEP is utilized to set up the 5-HT_{2C} agonist compounds QSAR model based on the descriptors calculated from the molecular structures by the software CODESSA. In order to learn the different influence on the EC50 of each descriptor, we also built a linear model based on heuristic method (HM). The results shown that GEP predicted ability is better than HM. To our knowledge, GEP method is firstly used for predicting the EC_{50} of the 5-HT_{2C} agonist compounds on the basis of the molecular structural descriptors.

2. Data and Method

2.1. Data

The structures and EC50 values for 5-HT_{2C} agonist were collected from the literature [8] and are listed in Table 1. Functional screenings were carried out in HEK293E cells expressing the human 5-HT_{2C} receptor. This research takes $LogEC_{50}$ as indicators of 5-HT_{2C} agonist. 39 compounds are divided into 25 compounds of training set and 13 compounds of test set randomly. We build QSAR model according to training set, and then use this model to predict the value of test set.

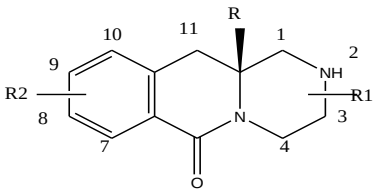
2.2. Calculation of the descriptors

In QSAR model, chemical structures are always expressed as numerical physical and chemical parameters (often called molecular descriptors) that calculated by professional computational chemistry software. The molecular descriptors are the most important factors showing the chemical information and affecting the quality of the QSAR model. All 5-HT_{2C} agonists were built molecular structural formula in Chemdraw software. Firstly, the geometry optimization was operated by the Hyperchem software, the calculation of molecular mechanics was with MM+. To get more accurate optimization, compounds were

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

further optimized with semi-empirical PM3 method. Then, new generated files would be precisely calculated in the MOPAC. Finally, the HIN files generated by geometrical optimization, the MNO files generated by MOPAC and the $logEC_{50}$ values would be put in CODESSA software to calculate the optimum molecular descriptors and build HM model. The molecular descriptors can be divided into six types: Constitutional descriptors, Geometrical descriptors, Electrostatic descriptors, Thermodynamic descriptors, CPSA Descriptors, and Topological descriptors. The several methods combined for quantifying the structural information about the molecule with advanced statistical analysis can satisfy the need to establish molecular structure-activity relationships in CODESSA.

Table 1 Experimental and calculated LogEC50 of 5-HT_{2C} agonists



compounds	R	R1	R2	LogEC50	GEP		HM	
					Pre.	Resid.	Pre.	Resid.
1	H	H	8-Cl	1.000	1.084	-0.084	1.361	0.361
2	H	H	9-Me	1.740	2.001	-0.261	2.068	0.328
3	H	2-Me	H	2.286	1.800	0.486	1.930	-0.356
4	H	H	H	1.556	1.470	0.086	1.571	0.015
5*	H	H	7-Me	0.699	0.728	-0.029	1.063	0.364
6	H	H	7-OMe	1.079	1.014	0.065	0.896	-0.183
7*	H	H	10-CF_3	3.523	3.382	0.141	4.032	0.509
8	H	H	8-CN	2.778	2.796	-0.018	2.715	-0.063
9	H	3-β-Me	H	3.242	2.303	0.939	2.198	-1.044
10*	H	H	11-β- Me	1.748	1.678	0.070	1.797	0.049
11	H	H	9-Pr	2.033	1.934	0.099	2.351	0.318
12*	H	H	9-OMe	2.382	1.898	0.484	2.068	-0.314
13	H	H	9-CN	2.785	2.515	0.270	2.562	-0.223
14	H	H	10-Me	1.447	1.884	-0.437	1.723	0.276
15	H	H	7-Me,9-Me	1.146	1.315	-0.169	1.089	-0.057
16*	Me	H	H	3.133	2.324	0.809	2.385	-0.749
17*	H	H	7-Et	1.146	1.268	-0.122	1.298	0.152
18	H	H	9-Me,10-OMe	1.886	2.141	-0.255	2.124	0.238
19	H	H	4-α-Me	0.477	1.209	-0.732	1.616	1.139
20*	H	H	9-Et	1.740	2.060	-0.320	2.113	0.373
21	H	H	10-OMe	1.079	1.159	-0.080	1.197	0.118
22*	H	H	10-OPr	1.845	1.246	0.599	1.613	-0.233
23	H	H	8-CF_3	1.255	1.630	-0.375	1.301	0.046
24	H	H	7-Cl	1.342	1.302	0.040	1.506	0.164
25	H	H	8-Me	0.903	0.915	-0.012	0.925	0.022
26	H	H	8-Et	0.602	0.841	-0.239	1.183	0.581
27	H	H	7-Et	1.681	1.844	-0.163	2.142	0.461
28*	H	H	8-OMe	0.477	0.939	-0.462	0.656	0.179
29	H	3-α-Me	H	2.519	2.303	0.216	2.198	-0.321
30*	H	H	7-cycPr	1.806	1.830	-0.024	1.698	-0.108
31	H	H	7-CF_3	1.204	1.010	0.194	0.813	-0.391
32	H	H	7-Me,10-OMe	0.778	0.993	-0.215	0.500	-0.278
33	H	4-β-Me	H	1.519	1.209	0.310	1.616	0.097
34*	H	H	7-CN	2.505	2.433	0.072	2.103	-0.402
35	H	H	9-CF_3	2.766	2.488	0.278	2.008	-0.758
36	H	H	9-Cl	2.233	2.416	-0.183	2.435	0.202
37	H	H	10-OEt	1.398	1.322	0.076	1.455	0.057
38*	H	H	7-Me,9-Et	1.964	1.109	0.855	1.400	-0.564
39*	H	H	9,10-$(CH2)_3O$-	1.204	1.323	-0.119	1.196	-0.008

The star '*' is test set

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

2.3. Heuristic method

HM can search all completely calculated descriptors, ultimately selected the best combination of descriptor. Before execution, HM pre filter the descriptors, well. The first is for descriptor linear restrictions. In HM algorithm, if the correlation coefficient between any two descriptors >0.80, the final selection will only choose one of them. Second to eliminate some descriptors, mainly is the parameters of partial compounds not have. T test value is less than a defined. HM will order descriptors in descending according to the correlation coefficient of parameter model. Each introduction descriptor is the biggest of all response correlation coefficient values. The performance of the model is mainly verified by R^2, R^2_{CV} and S^2. When R^2_{CV} increment is less than 0.02, the selection process terminated. R^2_{CV} is used to measure the reliability and robustness of QSAR model.

2.4. Gene expression programming

GEP, a non-parametric method, is invented by Ferreira in 1999, and it is an extension of genetic programming (GP). Gene expression programming is a genotype/phenotype system, in which a mathematical function defined as a linear chromosome with fixed length. The chromosomes are composed with multi-genes which are structurally organized in a head and a tail. The head of gene contains both function and terminal symbols, whereas the tail is terminal symbols only. GEP is individual chromosomes. Each chromosome is made up of one or more genes. Each gene is a fixed length string, divided into the head and tail. Head can also contain a terminator, and tail contains only the end operator [12]. Terminators are often constant. For a chromosome, the same length of each gene is fixed, so the length of each chromosome is fixed. The length of the body color is determined by the need to solve specific problems, but once we know head length, tail length can be calculated by formula. Suppose the head length of H, tail length of T, the relationship between t and h as: T = H*(n-1) + 1. N represents the function collection of input parameters that are the most number of parameters in the function. The function of chromosomes is subjected to modification by means of mutation, transposition, etc, to find a best solution in the form of expression trees (ETs).

The genetic code of GEP is very simple: a one-to-one relationship between the symbols of chromosome and the nodes they represent in the trees. Operation process of GEP method is first to create a random initial population, the population of each individual translated into K a expression, and calculate the fitness value to assess individual fitness. If there are not termination conditions, it will update population by genetic operations.

This process has been circulating, until meet the termination conditions. GEP includes single point of restructuring, recombination, mutation, insertion, genetic recombination and other genetic operations that have a specific meaning. The real answer from GEP is an expression, namely individual phenotype. The expression is obtained by chromosome. Despite the length of the chromosome is fixed, but due to the special encoding, expression length can be different. GEP genetic operators in addition to the selection operator and mutation operator of genetic algorithm, but also increased the conversion and insertion operator and restructuring grates. In order to keep the decode expression tree completely, the genetic operations need to keep the length of the gene and the tails end up with operators.

3. Results and Discussion

3.1. The HM model

354 descriptors were calculated by the CODESSA program totally for all the 5-HT_{2C} agonist compounds. HM can not only filter a group of molecular descriptors, but also build a linear model according to the descriptors. The values of R^2, R^2_{CV}, and the S^2 were used to judge the quality of the model. And R^2 and R^2_{CV} improved, S^2 decreased with the growing of number of molecular descriptors. It should be noted that, excessive molecular descriptors may cause over-parameterization. To avoid the possible multi-collinearity of molecular descriptor scales, Table 2 lists the correlation coefficients between the descriptors were selected.

Table 2 Correlation coefficients of the descriptors in HM model

	HACA-2	MERICA	MBOHA	RPCG	ACWP
HACA-2	1.000	0.056	0.130	0.184	0.221
MERICA		1.000	0.021	0.622	0.350
MBOHA			1.00	0.269	0.347
RPCG				1.000	0.751
ACWP					1.000

From the Table 2 we can see that there is any correlation coefficient more than 0.8. Obviously, all the descriptors are strongly orthogonal and the HM model has a pretty statistical reliability for predicting the $LogEC_{50}$ of the 5-HT_{2C} agonist. Taking the increase of the R^2 values of less than 0.02 should been chosen as the breakpoint criterion into account, 5 descriptors are selected in this study at last (Figure 1). The details

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

of five descriptors are given in Table 3. The plot of HM is shown in Figure 2. The QSAR model of $LogEC_{50}$ is built by HM as follows:

$$LogEC_{50} = 87.924 + 41.499HACA\text{-}2 + 411.89MERICA + 93.178MBOHA - 52.152RPCG - 48.339ACWP \quad (1)$$

$R^2 = 0.73$, $R^2_{cv} = 0.63$, $S^2 = 0.195$, F= 17.510

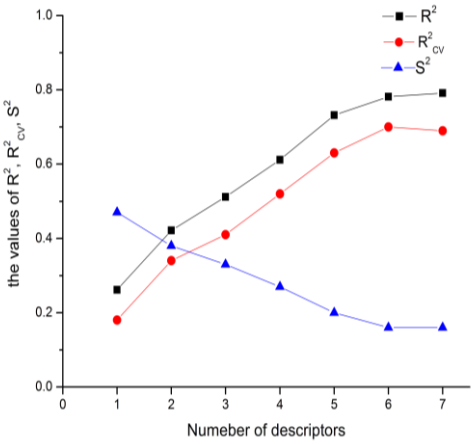


Figure 1. Influence of the number of descriptors on the R^2, R^2_{CV}, S^2.

Through the study of the descriptors, it may be possible to obtain some insight into factors that affect the activity of the compounds and understand the mechanism of 5-HT_{2C} agonist in treatment of obesity. In this study, five descriptors involved in the model belong to two types.

HACA-2/TMSA [Quantum-Chemical PC] (HACA-2) represents the area-weighted surface charge of hydrogen bonding acceptor atoms. The empirical partial charges in the molecule are calculated using the approach proposed by Zefirov. TMSA is the total molecular surface area. HACA-2 can be calculated by eq. (2).

$$HDCA2 = \sum_{A} \frac{q_A \sqrt{S_A}}{\sqrt{S_{tot}}} \quad (2)$$

Where S_A is the solvent-accessible surface area of H-bonding acceptor atoms that is selected by threshold charge, q_A is the partial charge on H-bonding acceptor atoms that is selected by threshold charge, and S_{tot} is the total solvent-accessible molecular surface area. HDCA is a basic descriptor of the compounds. When the structure can provide solvent-accessible molecular surface area, it will promote the distribution of the drug in the body, and will be conductive to the value of EC_{50} [13].

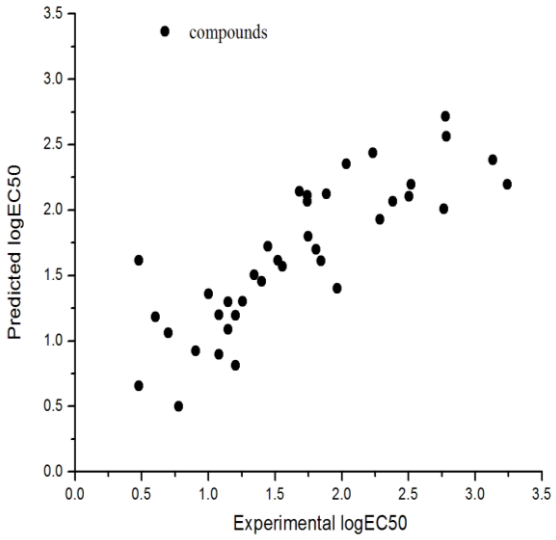


Figure 2. Plot of experimental and predicted $LogEC_{50}$ by HM.

Max electroph react index for a C atom (MERICA), as a quantum-chemical descriptor, provides a useful means of characterizing them charge transfer interactions between the repellent and the receptor. Thus the properties of charge play a key role in the compounds. It can be calculated by eq. (3):

$$E_A = \sum_{j \in A} \frac{C^2_{jLUMO}}{\varepsilon_{LUMO} + 10} \quad (3)$$

Where: ε_{LUMO} the lowest unoccupied molecular orbital energy, C_{jLUMO} the lowest unoccupied molecular orbital MO coefficients. This descriptor has the largest positive coefficients. It implies that the high max electroph react index for a C atom play an obstacle to efficacy of 5-HT_{2C} agonist [14].

Table 3 The corresponding physicochemical meaning and statistical parameters

Symbol	Physical-chemical meaning	Coefficient	Standard error	T-test
HACA-2	HACA-2/TMSA [Quantum-Chemical PC]	41.499	7.786	5.329
MERICA	Max electroph react index for a C atom	411.890	60.491	6.809
MBOHA	Min (>0.1) bond order of a H atom	93.178	28.559	3.262
RPCG	Relative positive charge (QMPOS/QTPLUS) [Quantum-Chemical PC]	- 52.152	9.4226	-5.534
ACWP	atomic charge weighted PPSA [Zefirov's PC]	-48.339	0.126	-3.813

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

Min (>0.1) bond order of a H atom (MBOHA) belongs to topological descriptors. It relates to the strength of intramolecular bonding interactions, the characteristic of the molecular stability, and their conformational flexibility as well as other valency-related properties [15]. The increasing of MBOHA can contribute to an increase of EC50. The bond order can affect the stability of compounds, and the high bond order compounds are not easy to be oxidized in the body. Thus, the metabolic process of the drug will slow down and the values of EC50 will increase.

Relative positive charge (QMPOS/QTPLUS) [Quantum-Chemical PC] (RPCG) and atomic charge weighted PPSA [Zefirov's PC] (ACWP) belong to CPSA Descriptors. CPSA descriptors can provide the information about the features of molecules responsible for polar intermolecular interactions. Most recently, Professional researchers got more details about the CPSA descriptors, evaluating their characteristics with regard to conformational dependence, sources of partial atomic charges, utility of whole molecule and substructure varieties, and the inclusion or exclusion of explicit hydrogens [16]. These two descriptors indicate the importance of the intramolecular electronic effects on the intermolecular multipole electrostatic interactions of a molecule in determining the EC50 of the 5-HT_{2C} agonist. Due to the nagetive sign of the regression coefficient, the lower the descriptor value the higher its agonistic effect.

3.2. The GEP model

A new method to generate the QSAR model could be provided by the software automatic problem solver (APS), which allows the easy optimization of intermediate solutions and the easy testing of the evolved models against a test set. GEP model also adopts the same molecular descriptors calculated by CODESSA. As a result, the 89th generation is a very good solution, with a R^2 of 0.82 for the training set and a R^2 of 0.78 for the test set. The C^{++} function is converted into corresponding equation:

$$LogEC50 = \sin\left(\tan\left(\left(\frac{x_2}{x_3}+x_4\right)\times\frac{x_3}{x_2}\right)\right)+\frac{x_3}{x_4-x_2}\times\sqrt{\tan(\tan x_2)}+\tan x_1+$$

$$\tan(\tan(\tan(\tan(\tan(\tan(\tan(\sin(\sqrt{x_5})))))))))+\sin\left(\sin\left(\sin\left(\frac{Logx_3}{\tan\frac{x_1}{x_2}}\right)\right)\right)+ \quad (4)$$

$$\sin\left((\tan x_5-1)\times\left(\frac{x_3}{x_4}-x_2\right)\right)\times x_4$$

In eq. (4), x_1 to x_5, represent the variables of HACA-2, MERICA, MBOHA, RPCG, ACWP, respectively. A good solution with R^2 of 0.82 and S^2 of 0.10 for training set, R^2 of 0.78 and S^2 of 0.18 for test training set (Figure 3, Figure 4) in GEP model. Compared with linear HM model, nonlinear GEP model can provide a better result.

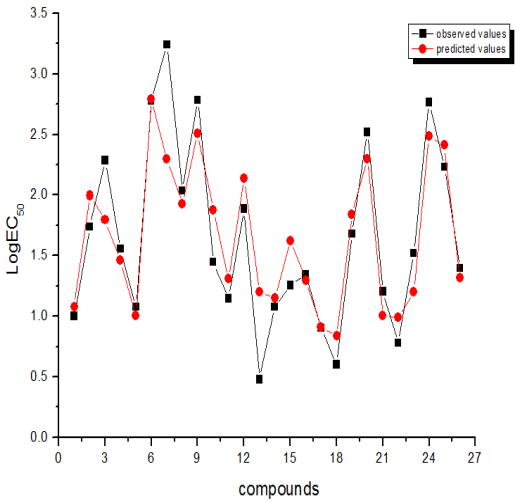


Figure 3. Fitting curve of experimental and predicted values of training set by GEP.

HM can quickly filter parameters to establish a stable linear QSAR model. However, the absorption, distribution, metabolism, and excretion of 5-HT_{2C} agonist are tanglesome. In many cases, the nonlinear relationship could get more satisfactory results than linear. The selection of five descriptors, which should be the most important physical chemistry properties for the structures, are core stages for the construction of QSAR model and a good prediction of EC50. Descriptors don't only meet the needs of the model, but also to have certain biological significance.

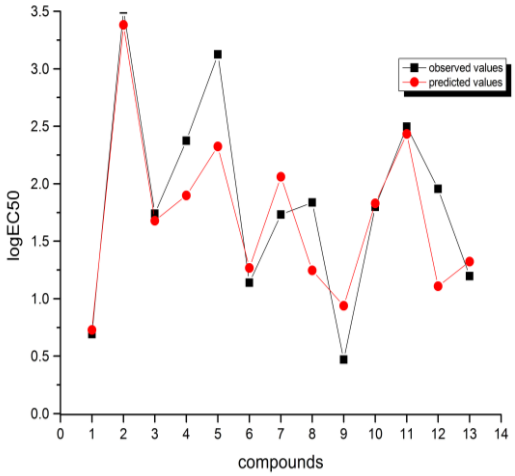


Figure 4. Fitting curve of experimental and predicted values of test set by GEP.

4. Conclusion

QSAR approach was successfully used to predict the activity of the 5-HT_{2C} agonist and achieved amazing results. The research shows that GEP is a pretty promising tool for building nonlinear approximation, the predictive power of GEP is better than HM. The adjustable chromosome number and the generation of

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

ETs enable the GEP to considerably stand out from current algorithms. The selection of molecular physics-chemistry parameters and the application of appropriate algorithms are the keys to QSAR analysis. At the same time, the theoretical model can provide instructions about internal mechanism of drugs and new ideas for further designing of novel 5-HT_{2C} agonist.

References

[1] K. Lv, Q.S. Tang, X.L. Li, X.K. Yang, M.J. Zhu, Journal of Beijing University of Traditional Chinese Medicine, 36 (2013) 327-332.

[2] X.M. Shen, Y.H. Du, T. Li, X.M. Shi, Zhenci Yanjiu, 29 (2004) 35-38.

[3] D. Stynen, N.W.H. Jansen, I.J. Kant, Journal of Psychosomatic Research, 79 (2015) 604-613.

[4] A. Duarte, J. Walker, S. Walker, G. Richardson, C. Holm Hansen, P. Martin, G. Murray, M. Sculpher, M. Sharpe, Journal of Psychosomatic Research, 79 (2015) 465-470.

[5] Q. Zhao, Y. Tang, S. Chen, Y. Lyu, A. Curtin, J. Wang, J. Sun, S. Tong, Neurophysiologie Clinique/Clinical Neurophysiology, 45 (2015) 435-443.

[6] J.E. Maurer-Spure, C. Pitendreigh, S. Misri, J. Psychiatry Neurosci, 32 (2007) 23-29.

[7] P.M. Pitychoutis, C. Dalla, A.C. Sideris, P.A. Tsonis, Z. Papadopoulou-Daifoti, Neuroscience, 210 (2012) 152-167.

[8] G.H. Zhao, C. Kwon, S.N. Bisaha, P.D. Stein, K. A. Rossi, X.Y. Cao, Bioorganic & Medicinal Chemistry Letters, 23 (2013) 3914-3919.

[9] N. Adhikari, A.K. Halder, A. Saha, K.D. Saha, T. Jha, Toxicology in Vitro, 29 (2015) 1392-1404..

[10] R.P. Verma, C. Hansch, European Journal of Medicinal Chemistry, 45 (2010) 1470-1477.

[11] J.T. Yuan, Y.P. Pu, L.H. Yin, Ecotoxicology and Environmental Safety, 84 (2012) 282-292.

[12] R.C. Li, X.X. Chen, F. Luan, Journal of Computational Science & Engineering, 10 (2014) 391-395.

[13] H.Z. Si, S.P. Yuan, K.J. Zhang, A.P. Fu, Y.B. Duan, Z.D. Hu, Chemometrics and Intelligent Laboratory Systems, 90 (2008) 15-24.

[14] J. Song, Z.D. Wang, Z.J. Han, J.Z. Chen, Z.K. Jiang, W.Q. Zheng, Z.B. Shang, Forest Products Journal, 32 (2012) 3-8.

[15] A.B. Sannigrahi, Advances in Quantum Chemistry, 23 (1992) 301-351.

[16] L.H. Cui, C.C. Jiao, Y.C. Liu, H.Z. Si, Cancer Cell Research, 1 (2014) 96-99.

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