Organic and Medicinal Chemistry
International Journal
ISSN 2474-7610

Research Article
Volume 4 Issue 3 - November 2017
DOI: 10.19080/OMCIJ.2017.04.555638

Organic & Medicinal Chem IJ

In silico Modelling of 2D, 3D Molecular Descriptors for Prediction Of Anticancer Activities Of Luteolin And Daidzin From Plants *Perilla ocymoides L and Glucine max L*

Pham Van Tat[3]*, Bui Thi Phuong Thuy[1], Tran Duong[2], Phung Van Trung[4], Hoang Thi Kim Dung[4] and Pham Nu Ngoc Han[3]

[1]*Faculty of Chemistry, Hue University of Science, Asia*

[2]*Faculty of Chemistry, Hue University of Education, Asia*

[3]*Faculty of Science and Engineering, Hoa Sen University, Asia*

[4]*Institute of Chemical Technology, Vietnam Academy of Science and Technology, Asia*

Submission: October 10, 2017; **Published:** November 20, 2017

***Corresponding author:** Pham Van Tat, Faculty of Science and Engineering, Hoa Sen University, Asia, Email: vantat@gmail.com

Abstract

Recently, we have isolated two flavonoids luteolin and daidzin from leaves of *Perilla ocymoides L* and *Glucine max L* in Viet Nam [1], with cytotoxic activity relatively strong in Hela cell line. To clarify the important nature of the relationships between structure and activity, the QSAR studies on Hela cell line incorporated the principal component analysis (PCA) technique and the artificial neural network (ANN) to construct the $QSAR_{PCA-ANN}$ relationships. The best multiple linear model $QSAR_{MLR}$ (with k = 6) values R^2_{train} of 0.854 and R^2_{pred} of 0.812, and QSARPCR (with k = 6) values R^2_{train} of 0.937 and R^2_{pred} of 0.889 were found by using the multiple linear regression technique. The artificial neural network $QSAR_{PCA-ANN}$ with architectural style I (6)-HL (9)-O (1) represented the values R^2_{train} of 0.993 and R^2_{pred} of 0.971. In the case the incorporated model $QSAR_{PCA-ANN}$ with the architecture I (6)-HL (9)-O (1) was exhibited the higher training and predicted quality. The anticancer activities of test substances resulting from those models are in good agreement with those from literature. The anti-cancer activities of two compounds luteolin and daidzin from leaves of *Perilla ocymoides L* and *Glucine max L* resulting from those models turn out to be agreement with experimental data.

Keywords: $QSAR_{MLR}$, $QSAR_{PCR}$ and $QSAR_{PCA-ANN}$ Model; Anti Cancer Activities Hela Cell Line

Abbreviations: PCA - Principal Component Analysis; ANN - Artificial Neural Network; QSAR - Quantitative Structure-Activity Relationship; PCR - Principal Component Regression; PCs - Principal Components; SE - Standard Error; LOO - Leave-One-Out; HMBC - Heteronuclear Multiple-Bond Correlation; HSQC - Heteronuclear Single-Quantum Correlation spectroscopy; QSAR - Quantitative Structure Activity Relationship; MLR - Multiple Linear Regression

Introduction

Natural products from plants are of interest in searching for new anti-cancer drugs and can have a direct effect on HeLa cell line and reduce side effects. Recently, we have isolated a few flavonoids from *Perilla ocymoides L and Glucine max L* [1] and tested *in vitro* activities pointed out the relatively strong impacts for cancer cells HeLa [2]. Flavonoids are polyphenolic compounds in most plants [3-5]. The flavonoids from *Perilla ocymoides L and Glucine max L* were also tested the biological activities in some different cancer cells. The flavonoids presented their activities and role of food within flavonoids in the cancer inhibition are widely studied [6-8]. In recent years, the computational methods are applied widely for the study of chemical properties and designing new drugs. The field of new drug design by *in sillico* method has become an important tool nowadays. In sillico study on quantitative relationships between structure and activity (QSAR) of natural products is concerned with the new drug researchers and pharmaceutical manufacturing facilitators. In Viet Nam, there are also a number of works of scientists from universities and institutions published in Viet Nam journals [9-11]. In the previous studies of 3-aminoflavonoid substances they have focused on the use of semi-empirical calculation [11]. Those studies showed an effect way for designing new drugs with the assisted computers. The In sillicao model it can be used to predict the biological activity of new drugs from the atomic charges and molecular descriptors. This method allows for the identification of an active-central location of molecule.

The set of flavones and isoflavones is known to have an important activity against cervical cancer cells [12-14]. This flavonoid group is also interested currently for researching in different directions such as the synthesis and metabolizing of natural products isolating them from plant [1]. The in sillico model of quantitative relationship between the structure of flavones and isoflavones and anticancer activity is an important issue for searching the flavonoid derivatives to be valid way. In this work, we report in the present paper the use of semi-empirical quantum calculations and construction of quantitative structure activity relationship (QSAR) models using 32 flavone and isoflavone derivatives [15]. The geometries of flavones and isoflavones are optimized by means of molecular mechanics (MM+). The 2D and 3D molecular descriptors resulting from geometric calculation are used to establish the multivariate models such as multiple linear regression (MLR), principal component regression (PCR) and artificial neural network (ANN). The anti-cancer activities GI50/□M of flavones and isoflavones in test group and two new flavonoids luteolin and daidzin from Perilla ocymoides L and Glucine max L [1] resulting from in sillico models are compared with those from experimental data..

Materials And Methods

a. Materials: To ensure an accurate capability of QSAR model, the dataset used for building and validating QSAR models consists of 32 compounds with anti-cancer activities $GI_{50}/\mu M$ for Hela cell line (GI_{50} is the concentration for 50% of maximal inhibition of cell proliferation) were reported by Wang et al. in the literature [2], as pointed out in (Figure 1). The value $\log GI_{50}$ is the subsequent dependent variable that defines the biological parameter for QSAR model.

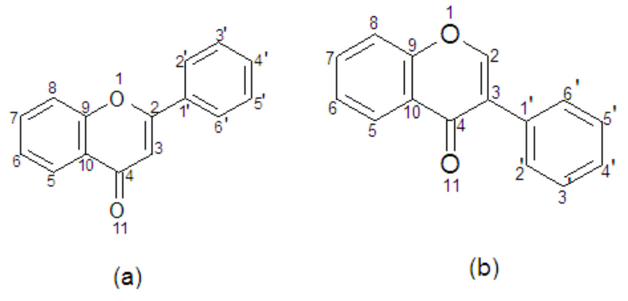


Figure 1: Molecular skeleton: a) flavone and b) isoflavone.

$$pGI_{50} = -\log GI_{50} \qquad (1)$$

Quantitative Structure-Activity Relationship (QSAR) studies have often been used to find correlations between biological activities and 2D and 3D molecular descriptors for compounds. We used the flavones and isoflavones reported by Wang et al. to calculate 2D and 3D molecular descriptors. The molecular descriptors are calculated with QSARIS program [16]. The multiple linear regression (QSARMLR) and principal component regression (QSARPCR) models are constructed with XLSTAT 2014 [17]. Because of the artificial neural networks are an artificial intelligent systems, they use a large number of interrelated data-processing neurons to emulate the function of brain. So the artificial neural network (QSARPCA-ANN) can be constructed with program Visual Gene Developer 1.7 [18].

b. Constructing QSAR models: Linear regression is without doubt the most frequently used statistical method. The multiple regression (several explanatory variables) and simple linear regression are identical linear regression methods in the overall concept as well as calculation techniques. The principle of linear regression is to model a quantitative dependent variable Y though a linear combination of k quantitative explanatory variables, x1, x2, ..., xk. In the case where there are N observations, the estimation of the predicted value of the dependent variable Y is given by [17,19]:

$$\mathrm{Y} = \sum_{i=1}^{k} a_i x_i + b \quad (2)$$

Where Y is the experimental activity pGI_{50},exp, x_i is kth molecular descriptor.

Values R^2_{train} and R^2_{pred} are calculated by

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\mathrm{Y}_i - \hat{\mathrm{Y}}_i)^2}{\sum_{i=1}^{N} (\mathrm{Y}_i - \bar{\mathrm{Y}})^2} \qquad (3)$$

Here Y*i*, $\hat{\mathrm{Y}}_i$ are experimental pGI_{50}, $_{exp}$ and predicted pGI_{50}, pred value; $\bar{\mathrm{Y}}$ is mean of experimental values.

The predicted results derived from the QSAR models are validated and compared with experimental data base on the relative errors (ARE, %) as [3,7]:

$$ARE,\% = 100\left|(pGI_{50,exp} - pGI_{50,pred})/pGI_{50,exp}\right| \qquad (4)$$

The average value of absolute relative errors ARE, % [6] is calculated and used for assessing the global uncertainty of QSAR model

$$MARE,\% = \frac{100}{N}\left|\frac{(pGI_{50,exp} - pGI_{50,pred})}{pGI_{50,exp}}\right| \qquad (5)$$

With *N* is number of activity values.

Principal component regression (PCR) model [17,20] is a regression technique using principal component analysis (PCA) when evaluating regression coefficients. PCR presents a technique for finding structure in datasets. Its object is to group correlated variables, replacing the earlier descriptors by new set called principal components (PCs). These PC's are uncorrelated and are developed as a simple linear aggregation of earlier variables. It moves the data into a new set of axes such that first few axes indicate most of the variations within the data. First PC (PC1) is expressed in the direction of maximum variance of the whole dataset. Second PC (PC2) is the direction that defines the maximum variance in orthogonal subspace to PC1. Consequent components are taken orthogonal to the particular formerly

How to cite this article: Pham V T, Bui T P T, Tran D, Phung V T, Hoang T K D, et al. *In silico* Modelling of 2D, 3D Molecular Descriptors for Prediction Of Anticancer Activities Of Luteolin And Daidzin From Plants *Perilla ocymoides L* and *Glucine max L*. Organic & Medicinal Chem IJ. 2017; 4(3): 555638. DOI: 10.19080/OMCIJ.2017.04.555638

chosen and defines best of remaining variance, by locating the data on new set of axes, it can points major fundamental structures certainly. Value of each point, when moved to a given axis, is called the PC value. PCA chooses a new set of axes for the data. These are chosen in decreasing order of variance within the data. The aim of principal component regression PCR is the computation of values of a response variable on the basis of chosen PCs of independent variables [16,17,20].

Results And Discussion

a. Calculation of molecular descriptors: The program HyperChem 8.05 [21] was used for designing the flavonoid molecules. The molecular structures were optimized by means of MM+ molecular mechanics. The molecular descriptors of molecules were calculated by computational techniques of QSARIS [16] using the optimized geometries. The molecular descriptors were used to construct the multiple linear regression ($QSAR^{MLR}$), principal component regression ($QSAR^{PCR}$) and artificial neural network ($QSAR_{PCA-ANN}$) model [4,5].

b. Development of $QSAR_{MLR}$ model: Before conducting the $QSAR_{MLR}$ model, the activity values GI_{50} (μM) are transformed into the values pGI_{50} to adapt the statistical properties. The values pGI_{50} (μM) are most appropriate value for modelling the relationships between molecular descriptors and activities. The $QSAR_{MLR}$ models were established by using the relationship of the geometric predictors and biological activities pGI_{50} [16]. The QSAR models in this work obtained by two different approaches: (i) cases are selected randomly for training set, and (ii) remaining cases for validation of predictability (test set). There are several methods for selecting the training set. The simplest way is the random selection. In this work, the original data is divided into training and validation set. The accurate predictability of QSAR model is evaluated by comparing the predicted and observed activities of the substances in test set without training set.

Table 1: Molecular structure and activities GI_{50} (μM) of flavones and isoflavones [2].

Substance	Name	Substitutive site	Substitutes R	GI50 (μM)	
Training set for establishing QSAR models					
Fla1	1a -1	flavone	C_3	$-OCH_2CCH_3=NOH$	2.0
Fla2	1b	flavone	C_6	$-OCH_2CCH_3=NOH$	1.2
Fla5	4a	flavone	C_3	$-OCH_2CCH_3=NOCH_3$	2.0
Fla6	5a	flavone	C_3	$-OCH_2CCH_3=NOCH_3$	0.9
Fla7	6a	flavone	C_7	$-OCH_2CCH_3=NOCH_3$	2.2
Isofla8	7a	isoflavone	C_7	$-OCH_2CCH_3=NOCH_3$	8.5
Fla10	8a	flavone	C_3	$-OCH_2C(4\text{-}F\text{-}C6H4)=NOH$	2.1
Fla11	9a	flavone	C_3	$-OCH_2C(4\text{-}CH3O\text{-}C6H4)=NOH$	2.0
Fla12	3b	flavone	C_6	$-OCH_2C(C_6H_5)=NOH$	0.8
Fla13	10a	flavone	C_6	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOH$	1.6
Fla14	11a	flavone	C_3	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOH$	1.0
Fla15	4b	flavone	C_7	$-OCH_2C(C_6H_5)=NOH$	2.0
Fla16	5b	flavone	C_7	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOH$	2.0
Fla17	12a	flavone	C_7	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOH$	2.0
Isofla18	13a	isoflavone	C_7	$-OCH_2C(C_6H_5)=NOH$	9.0
Isofla19	14a	isoflavone	C_7	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOH$	7.8
Isofla20	15a	isoflavone	C_7	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOH$	7.6
Fla21	16a	flavone	C_3	$-OCH_2C(C_6H_5)=NOCH_3$	1.6
Fla23	18a	flavone	C_3	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOCH_3$	2.0
Fla24	19a	flavone	C_6	$-OCH_2C(C_6H_5)=NOCH_3$	2.4
Fla25	20a	flavone	C_6	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOCH_3$	2.3
Fla26	21a	flavone	C_6	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOCH_3$	2.0
Fla27	22a	flavone	C_7	$-OCH_2C(C_6H_5)=NOCH_3$	6.6
Fla28	23a	flavone	C_7	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOCH_3$	2.7
Fla29	24a	flavone	C_7	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOCH_3$	2.5
Isofla30	25a	isoflavone	C_7	$-OCH_2C(C_6H_5)=NOCH_3$	8.2
Isofla31	26a	isoflavone	C_7	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOCH_3$	6.4
Test set for validating QSAR models					
Fla3	2a-3	flavone	C_7	$-OCH_2CCH_3=NOH$	2.0

Isofla4	3a-4	isoflavone	C_7	$-OCH_2CCH_3=NOH$	9.8
Fla9	2b	flavone	C_3	$-OCH_2C(C_6H_5)=NOH$	1.8
Fla22	17a	flavone	C_3	$-OCH_2C(4\text{-}F\text{-}C_6H_4)=NOCH_3$	2.0
Fla28	23a	flavone	C_7	$-OCH_2C(4\text{-}F\text{-}C6H_4)=NOCH_3$	2.7
Isofla32	6b	isoflavone	C_7	$-OCH_2C(4\text{-}CH_3O\text{-}C_6H_4)=NOCH_3$	7.3

Table 2: The QSARMLR models (k from 2 to 10) with change of values R2train, R2pred and SE.

k	The 2D, 3D molecular descriptors in models	R^2_{train}	R^2_{pred}	SE
2	LogP, MaxNeg	0.756	0.731	0.430
3	LogP, MaxNeg, ka2	0.774	0.732	0.417
4	LogP, MaxNeg, ka2, SdO	0.805	0.772	0.390
5	LogP, MaxNeg, MaxQp, SdO, ka2	0.832	0.756	0.365
6	LogP, MaxQp, Ovality, SdO, SdssC, ka3	0.854	0.812	0.342
7	LogP, MaxNeg, MaxQp, ka2, SdO, ABSQ, ABSQon,	0.836	0.721	0.365
8	LogP , MaxNeg, MaxQp, ka2, SdO , ka3, ABSQ, ABSQon	0.837	0.693	0.367
9	LogP, MaxNeg, MaxQp, ka2, SdO, ka3, ABSQ ,ABSQon, Dipole	0.838	0.682	0.369
10	LogP, MaxNeg, MaxQp, ka2, SdO, ka3, ABSQ, ABSQon, Dipole, Ovality	0.841	0.650	0.368

In recent years theoretical and experimental researchers have focused an increasing attention on finding the most efficient tools for selecting molecular descriptors in QSAR studies [3,4,10]. The change of values R_2^{train}, R^2_{pred} and SE (standard error) in the $QSAR_{MLR}$ models with the 2D and 3D predictors, respectively are pointed out in (Table 1). To have those $QSAR_{MLR}$ models, the 2D and 3D molecular descriptors were selected by using forward and backward algorithm. The selection process for 2D and 3D descriptors based on the change of the statistical values $R^2_{train,}$ SE and R^2pred. The values R2pred of the $QSAR_{MLR}$ models were calculated by using the cross-validated technique with leave-one-out (LOO) method. The 9 fitness models are shown in (Table 2).

Table 3: Statistical values and valuable contribution percentages MPmxk,% and GMPmxk,% for 2D and 3D molecular descriptors in the models $QSAR_{MLR}$ (with k of 5, 6 and 7).

Variable	$QSAR_{MLR}$			MP_mx_k,%			GMP_mx_k,
x_k	k = 5	k = 6	k = 7	k = 5	k = 6	k = 7	%
R^2_{train}	0.832	0.854	0.836				
R^2_{adj}	0.820	0.841	0.820				
SE	0.365	0.342	0.365				
R^2_{pred}	0.756	0.812	0.721				
Constants	3.883	8.509	4.790				
ABSQ	-0.222	-	-0.257	27.945	18.636	19.005	21.862
ABSQon	-	-	0.0143	-	-	0.433	0.144
MaxQp	3.416	2.8540	3.588	24.043	25.862	25.908	25.271
MaxNeg	-	-	6.122	-	24.203	23.890	16.031
SdO	0.0125	0.0247	0.0126	6.192	3.792	3.665	4.550
ka2	0.133	-	0.143	27.484	17.617	17.426	20.842
LogP	0.156	0.2192	0.163	15.651	9.829	9.672	11.717
Ovality	-	-3.6969	-	5.292	4.393	2.315	4.000
SdssC	-	0.2969	-	4.382	5.613	7.236	5.744
ka3	-	0.3635	-	5.351	9.324	3.473	6.049

The $QSAR_{MLR}$ models (with k of 2 to 10) that were arranged in an orderly change of R^2_{train}, SE and R^2_{pred}, as given in (Table 2). The values of R^2_{train} and R^2_{pred} from $QSAR_{MLR}$ models (with k from 5 to 7) are higher than others. In particular, the $QSAR_{MLR}$ model (with k = 6) has given the highest values R^2_{train} of 0.854 and R2pred of 0.812. So, three best models (with k of 5, 6 and 7) are chosen to determine the significant contribution of 2D, 3D descriptors. The valuable contribution percentages $MP_{m}x_{k}$,% and $GMP_{m}x_{k}$,% [3], with the statistical parameters of three models (with k of 5, 6 and 7), respectively, are given in (Table 3).

The contribution percentages $MP_{m}x_{k}$, %, $GMP_{m}x_{k}$, % [3,7,11] of the models (with k of 5, 6 and 7), respectively are calculated by formula

$$MP_{m}x_{k},\% = \frac{1}{N}\sum_{j=1}^{N}\left(100.\left|b_{m,i}x_{m,i}\right|/\mathrm{C}_{\mathrm{total}}\right) \text{ with } \mathrm{C}_{\mathrm{total}} = \sum_{j=1}^{k}\left|b_{m,k}x_{m,k}\right| \qquad (6)$$

Where N the total number of cases, m number of variables. The global average contribution percentage $GMP_{m}x_{k}$, % of each independent variable for 3 models is determined by the formula

$$GMP_{m}x_{k},\% = \frac{1}{n}\sum_{n=1}^{3}MP_{m}x_{k} \qquad (7)$$

With n number of models

The contribution percentages $GMP_{m}x_{k}$, % in Table 3 depicted the important level of 2D and 3D molecular descriptors for flavonoid compounds. For the $QSAR_{MLR}$ models in Table 3 the important significance of 2D and 3D molecular descriptors is arranged by using values $GMP_{m}x_{k}$, %: MaxQp > ABSQ > ka2 > MaxNeg > LogP > ka3 > SdssC > SdO > Ovality > ABSQon. The molecular descriptors MaxQp, ABSQ, ka2, MaxNeg and LogP can be considered such as the most important contribution for each molecule. Besides these molecular descriptors exhibit by important nature of carbonyl group $C_4 = O_{11}$ and atom O_1. These atoms wear the free electron pair conjugating with π electronic bond $C_2 = C_3$, and $C_4 = O_{11}$ to form a conjugate system. The carbonyl group $C_4 = O_{11}$ exhibited fully reactive nature of carbonyl substance [2]. So, these descriptors can be demonstrated quantitatively total charges ABSQ, MaxQp and MaxNeg on molecule based on the values GMPmxk, % and these are also consistent with the verdicts from experimental evaluation [16,23]. Furthermore, the atomic positions C_6 and C_3' on molecule are the vacant positions and can be explored for attaching the new function groups [9,23,24]. The various atoms seem to be the important impacts for biological activities GI_{50}. So these sites are chosen for attaching the new substitutes to construct new flavonoids. Similarly the atom C_2' is also empty position and also can be utilized to attach the new function group. Those sites hope to constitute the new compounds with higher activity than sample compound. Also this way, the new flavonoids isolated from leaves of *Perilla ocymoides L* and *Glucine max L* are also used such as lead compounds to design new drugs. This is also showed in below discussion.

c. Development of $QSAR_{PCR}$ model: The molecular descriptors were applied to under goes principal component regression PCR technique to create $QSAR_{PCR}$ model with simulated anealining variable selection mode by using PCR model [17]. The best $QSAR_{MLR}$ model (with k = 6) is selected to generate the QSARPCR model [16,17]. The 6 independent variables MaxQp, SdO, ka3, LogP, Ovality and SdssC were carried out to analyse the principle components. The principle component regression $QSAR_{PCR}$ model is generated with 6 principle components which are corresponding to the original descriptors of $QSAR_{MLR}$ model (with k = 6), as exhibited in equation (8):

$$pGI_{50} = 5.48356 + 0.38027\times PC_1 - 0.11868\times PC_2 + 0.34789\times PC_3 + 0.06995\times PC_4 + 0.21850\times PC_5 + 0.35057\times PC_6 \qquad (8)$$

With n = 26; R^2_{train} = 0.937; R^2_{Adj} = 0.9106, R^2_{Pred} = 0.889, SE = 0.342, MSE = 0.1236; F = 63.172

The number of principle components is extracted by the principle component analysis technique and the the correlation between pGI_{50} and pGI_{50}, pred values is pointed out in (Figure 2).

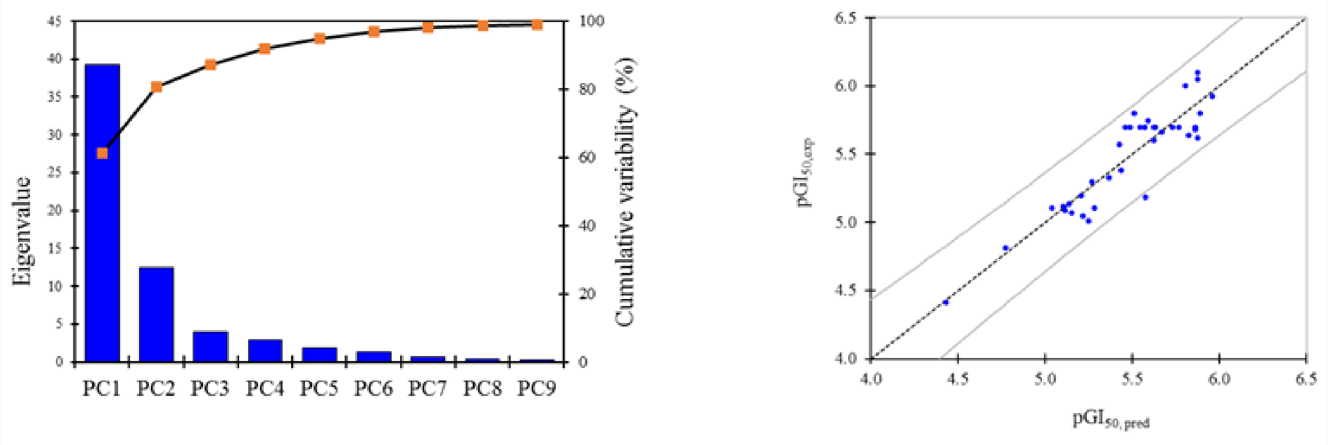


Figure 2: The nature of principle component and correlation between $pGI_{50},_{exp}$ and $pGI_{50},_{pred}$ values

How to cite this article: Pham V T, Bui T P T, Tran D, Phung V T, Hoang T K D, et al. *In silico* Modelling of 2D, 3D Molecular Descriptors for Prediction Of Anticancer Activities Of Luteolin And Daidzin From Plants *Perilla ocymoides L* and *Glucine max L*. Organic & Medicinal Chem IJ. 2017; 4(3): 555638. DOI: 10.19080/OMCIJ.2017.04.555638

d. **Building $QSA_{RPCA-ANN}$ model:** The $QSAR_{PCA-ANN}$ model is built by the neuro-fuzzy technique with the genetic algorithms using program Visual Gene Developer v1.7 [18]. The artificial neural network has an architecture style I(6)-HL(9)-O(1); it consists of input layer I(6) with 6 neurons such as 6 principle components in equation (8) PC_1, PC_2, PC_3, PC_4, PC_5 and PC6; the input neurons are corresponding to LogP, MaxQp, Ovality, SdO, SdssC and ka3; the neuron of output layer O(1) is the biological activity pGI_{50}; the hidden layer HL(9) consists of nine neurons. This neural network I(6)-HL(9)-O(1) used the back propagation algorithm to train the network.

The back propagation algorithm looks for the minimum of the error function in weight space using the method of gradient descent. The sigmoid function is used to transfer on each node of neural network; the training parameters of neural network are the training rate of 0.7 and learing rate of 0.7; the goal monitoring error MSE = 0.000816 with 10,000 iteration. After training the $QSAR_{PCA}$-ANN with architecture I(6)-HL(9)-O(1) pointed out the values R^2_{train} of 0.993 and R^2_{pred} of 0.971. But in the case the $QSAR_{PCR}$ model gave values R^2_{train} = 0.937 and R^2_{pred} = 0.889; and the $QSAR_{MLR}$ model (with k = 6) gave values R^2_{train} of 0.854 and R^2_{pred} of 0.812.

e.Isolation of luteolin and daidzin from plant

i. **Chemicals and equipment:** In this work, we used the chemicals and the equipments for isolating and purifying two flavonoids luteolin and daidzin before determining the substance structures by 1H-NMR and 13C-NMR spectrum [25]. The following materials are used to isolate the flavonoids in

ii. Silica gel with the particle size in range 0.04 to 0.06 mm was used for ordinary and Rp18 phase chromatography.

iii. Thin-layer chromatography was implemented by the thin plate DC-Alufolien F254 (Merck) for the ordinary phase and Rp18 F254s (Merck) for the reverse-phase chromatography.

iv. Solvents used for the isolation processes: hexane, petroleum ether, chloroform, methanol, ethyl acetate, ethanol, acetone, distilled water.

v. UV handheld lamps, 254 and 365 nm UVITEC effect.

vi. Vacuum Evaporators Buchi – 111 and Water Bath cooker JULABO 461.

vii. Infrared heating equipment SCHOTT.

viii. Chromatography column with diameter range 2 to 5.5 cm.

ix. Analytical Balances AND HR 200.

f. Isolation process of luteolin and daidzin: To isolate and purify the luteolin and daidzin compound from the leaves of Perilla ocymoides L and Glucine max L we used the techniques of thin-layer and column chromatography [25], as exhibited in (Figures 3-4). After isolating the compounds their structures were identified by the different spectrum as

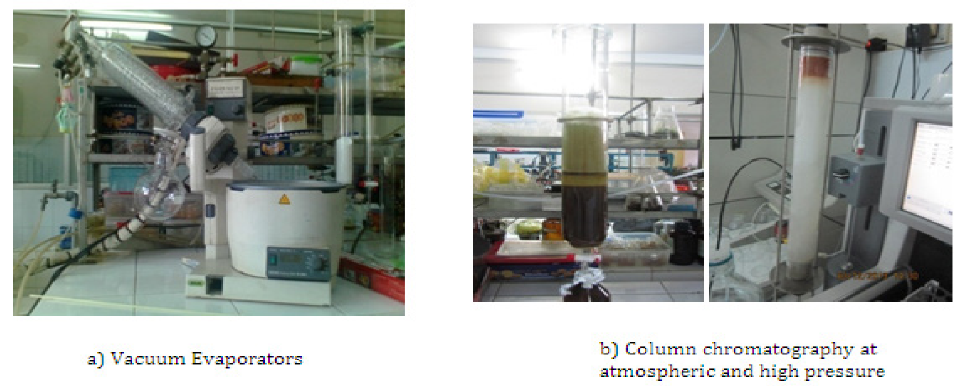


Figure 3: Separate equipment for two flavonoids luteolin and daidzin.

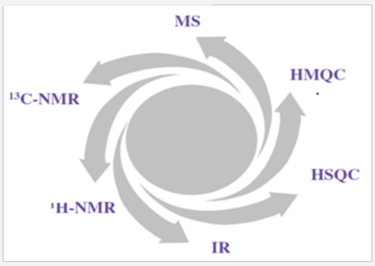


Figure 4: General diagram for identificating structures of luteolin and daidzin [24].

i. Melting temperature carried out on Electrothermal IA 9000 series, using unadjusted capillary.

ii. Column chromatography with silica gel for ordinary-phase, reverse-phase chromatography Rp 18 and sephadex techniques combined with thin-layer chromatography.

iii. Substances were detected by ultraviolet light at wavelengths 254 nm and 365 nm or reagent used is liquid H_2SO_4/EtOH or $FeCl_3$/EtOH.

iv. Nuclear magnetic resonance spectrum (NMR) ^{1}H-NMR

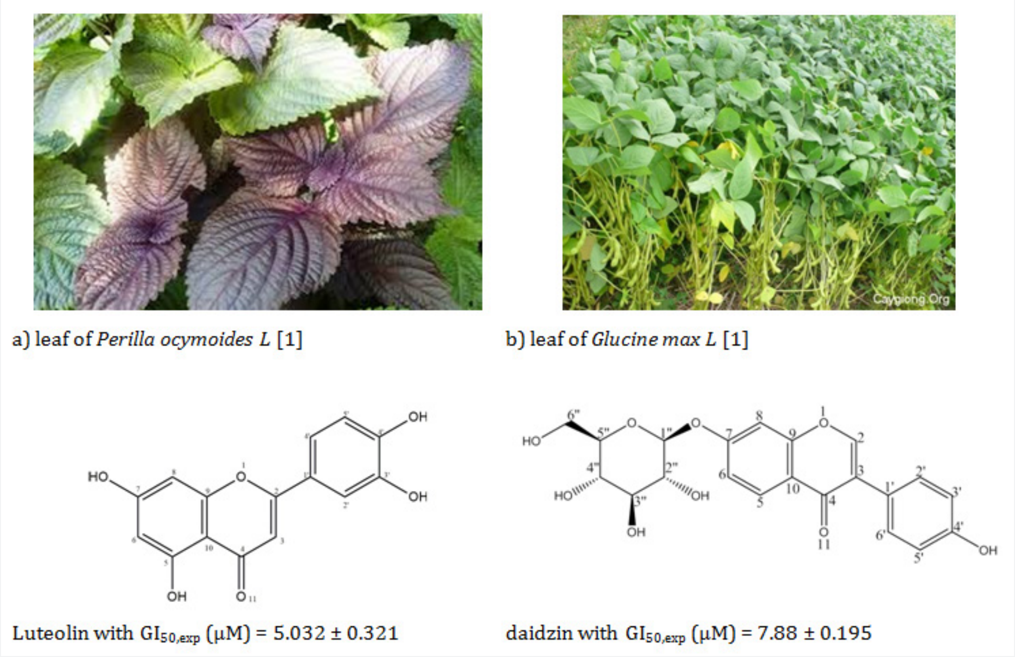
How to cite this article: Pham V T, Bui T P T, Tran D, Phung V T, Hoang T K D, et al. *In silico* Modelling of 2D, 3D Molecular Descriptors for Prediction Of Anticancer Activities Of Luteolin And Daidzin From Plants *Perilla ocymoides L* and *Glucine max L*. Organic & Medicinal Chem IJ. 2017; 4(3): 555638. DOI: 10.19080/OMCIJ.2017.04.555638

(500 MHz) and ^{13}C-NMR (125 MHz) implemented on Bruker AM500 FT-NMR Spectrometer.

g. Prediction of biological activity for new substances: The predictability of the constructed models QSAR$_{MLR}$, QSAR$_{PCR}$ and QSAR$_{PCA-ANN}$ was evaluated carefully by using the leave-one-out (LOO) technique to determine R$^2_{pred}$; the flavonoids in Table 1 were divided randomly into the training group of 26 compounds and the test group of 6 compounds. The anticancer activities pGI$_{50}$ of 6 flavonoids in the test group in Table 1 with 2 new flavonoids luteolin and daidzin isolated from the leaves of *Perilla ocymoides* L and *Glucine max L* [1] are predicted from those QSAR models. The predicted activities of 6 flavonoids in test group and new substances luteolin and daidzin resulting from QSAR models were compared to experimental data, as presented in (Table 4). For new substances luteolin and daidzin we carried out to test the in vitro activity on Hela cell line in laboratory of molecular biology of the genetic department at Ho Chi Minh University of science (Figure 5).

Table 4: Biological activities pGI$_{50}$ of 6 flavonoids in test group and two new substances luteolin and daidzin resulting from models QSAR$_{MLR}$, QSAR$_{PCR}$ and QSAR$_{PCA}$-ANN.

Substance	Ref.	pGI$_{50}$,$_{exp}$	pGI$_{50}$,$_{pred}$			ARE,%		
			QSAR$_{MLR}$	QSAR$_{PCR}$	QSAR$_{PCA-ANN}$	QSAR$_{MLR}$	QSAR$_{PCR}$	QSAR$_{PCA-ANN}$
Fla3	[8,12,13]	5.699	5.632	5.560	5.673	1.179	2.439	0.454
Isofla4	[8,12,13]	5.009	5.077	5.112	5.123	1.352	2.060	2.278
Fla9	[8,12,13]	5.745	5.690	5.740	5.687	0.954	0.082	1.006
Fla22	[8,12,13]	5.699	5.785	5.614	5.765	1.511	1.491	1.162
Fla28	[8,12,13]	5.569	5.653	5.559	5.668	1.510	0.172	1.783
Isofla32	[8,12,13]	5.137	5.088	5.115	5.064	0.948	0.422	1.413
luteolin	this work	5.032	4.415	4.979	5.106	12.262	1.053	1.476
daidzin	this work	5.103	4.592	4.637	4.982	10.014	9.124	2.382
					MARE,%	3.716	2.106	1.494



a) leaf of *Perilla ocymoides L* [1]

b) leaf of *Glucine max L* [1]

Luteolin with GI$_{50,exp}$ (μM) = 5.032 ± 0.321

daidzin with GI$_{50,exp}$ (μM) = 7.88 ± 0.195

Figure 5: The molecular structures of: a) luteolin and b) daidzin.

The luteolin structure was identified by using the different spectra such: ^{1}H-NMR (DMSO-d$_6$, 500 MHz, δ ppm) with HSQC, HMBC: δ 6.65 (1H; s, H$_3$); 6.19 (1H; d; J = 2Hz, H$_6$); 6.45 (1H; d; J = 2Hz, H$_8$); 7.4 (1H; s H$_2$'); 6.89 (1H; d; J = 8Hz, H$_{5'}$); 7.41 (1H; d; J = 8Hz, H$_{6'}$); 12.95 (1H, s; C5-OH); 9.4 (1H, s, C$_{4'}$-OH); 9.9(1H, s, C3'-OH); 10.84 (1H, s, C7-OH). The 13C-NMR spectrum was employed to have more information such as combining ^{13}C-NMR (DMSO-d$_6$, 125 Hz) with spectrum DEPT, HSQC, HMBC: δ163.1 (C$_2$); 102.8(C$_3$); 181.6(C$_4$); 161.4(C$_5$); 98.9(C$_6$); 164.1(C$_7$); 93.8(C$_8$); 157.3 (C$_9$);103.7(C$_{10}$); 121.5(C$_{1'}$); 113.4(C2'); 145.7(C$_{3'}$); 149.7(C4'); 116.0(C$_{5'}$); 118.9(C6'). Interaction of atom C and H in heteronuclear multiple-bond correlation (HMBC) and heteronuclear single-quantum correlation spectroscopy (HSQC) were pointed out the atomic sites: H$_6$- C$_5$- C$_7$- C$_8$-C$_{10}$; H$_8$- C$_6$- C$_7$-

How to cite this article: Pham V T, Bui T P T, Tran D, Phung V T, Hoang T K D, et al. *In silico* Modelling of 2D, 3D Molecular Descriptors for Prediction Of Anticancer Activities Of Luteolin And Daidzin From Plants *Perilla ocymoides L* and *Glucine max L*. Organic & Medicinal Chem IJ. 2017; 4(3): 555638. DOI: 10.19080/OMCIJ.2017.04.555638

C_9- C_{10}; $H_{2'}$- C_2- $C_{3'}$- $C_{4'}$- $C_{6'}$; $H_{3'}$-$C_{1'}$- $C_{2'}$- $C_{4'}$- $C_{5'}$; $H_{5'}$- $C_{1'}$- $C_{3'}$- $C_{4'}$- $C_{6'}$; $H_{6'}$- C_2- $C_{1'}$- $C_{2'}$- $C_{4'}$- $_{C5'}$.

For the substance daidzin the molecular structure was also identified by the spectrum ^{1}H-NMR: δ 8,06 (1H, d, J = 8,0Hz, H_5); δ 7,15 (1H, D, J = 1,5 Hz, H_6); δ 7,23 (1H, d, J = 2,0Hz, H_8); δ 6,84 ppm (2H, d, J = 6,5 Hz, H3', $H_{5'}$); δ 7,42 (2H, d, J = 8,0 Hz, $H_{6''}$, $H_{2'}$); Also, we used specrum 13C-NMR (DMSO-d_6, 125 Hz) with spectra DEPT, HSQC, HMBC: δ153,2 (C_2), δ 122,3 (C_3), δ 174,7 (C_4), δ 126,9 (C_5), δ 115,6 (C_6), δ 161,3 (C_7), δ 103,4 (C_8), δ 157,0 (C_9), δ 118,5 (C_{10}), δ 123,7 ($C_{1'}$), δ 130,0 ($C_{2'}$), δ 115,0 ($C_{3'}$), δ 157,2 ($C_{4'}$), δ 115,0 ($C_{5'}$), δ 130,0 ($C_{6'}$), δ 100,0 ($C_{1''}$), δ 73,1 ($C_{2''}$), δ 76,5 ($C_{3''}$), δ 69,6 ($C_{4''}$), δ 77,2 ($C_{5''}$), δ 60,6 ($C_{6''}$). The molecular structures of new substances luteolin and daidzin are shown in (Figure 5). The predicted activities from QSAR models were compared with experimental data and with each other upon the average value of absolute relative error MARE, %. The values MARE, % showed that the predictability of the model QSARMLR is lower than models QSARPCR and QSARPCA-ANN, as given in (Table 4). After using the QSAR models to predict the anticancer activities pGI_{50} of six flavonoids in test group and two new flavonoids luteolin and daidzin, the errors of QSAR models can be accepted in uncertainty range of experimental measurements. Consequently, the models $QSAR_{MLR}$, $QSAR_{PCR}$ and $QSAR_{PCA-ANN}$ exhibited in good adaptability for predicting the activities of new substances. In this work, we selected the new substance luteolin isolating from Perilla ocymoides L to design new substances. The new functional group are substituted at the vacant positions C_6, $C_{2'}$ and $C_{3'}$.

Table 5: The anticancer activities pGI_{50} of 5 new flavonoids resulting from QSARPCA-ANN model.

New substance	C_6	$C_{2'}$	$C_{3'}$	pGI50	Method in this work
luteolin	H	H	H	5.032	in vitro test on Hela
flav-1	H	NO_2	H	5.781	$QSAR_{PCA-ANN}$
flav-2	CH_3CO-	H	NO_2	6.155	$QSAR_{PCA-ANN}$
flav-3	CH_3CO-	NO_2	H	6.158	$QSAR_{PCA-ANN}$
flav-4	NO_2	NO_2	H	6.488	$QSAR_{PCA^-ANN}$
flav-5	NO_2	H	NO_2	6.538	$QSAR_{PCA-ANN}$

The substance luteolin was used such as lead compound for designing 5 new various compounds. The positions C_6, C_2' and $C_{3'}$ were substituted the new functional groups; and the biological activities pGI_{50} of the new designed flavonoids were predicted by using $QSAR_{PCA-ANN}$ model, as given in (Table 5). The predicted results pGI_5 for 5 new designed substances are compared with experimental activity of luteolin, as depicted in (Figure 6). The activity GI_{50} (μM) of five new designed compounds from luteolin by substituting new functional groups into C_6, $C_{2'}$ and $C_{3'}$ sites are stronger than lead compound luteolin. Herein the new designed compounds will promise to forward a designing plan for the new pharmaceutical products from natural products.

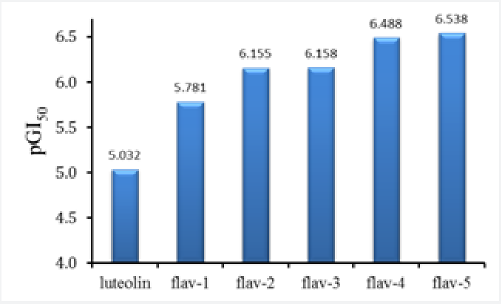


Figure 6: Comparison between values pGI_{50} of five new flavonoids with luteolin.

Conclusion

The use of computational methods constructed successfully the *in sillico* models with relationships between the 2D, 3D molecular descriptors and anti-cancer activities GI50 (μM) of flavonoids. The $QSAR_{MLR}$ model showed the important contribution descriptors MaxQp, SdO, ka3, LogP, Ovality and SdssC on flavonoids which effect an in vitro activity on Hela cell line. The *in sillico* model QSAR also found out helpfully the most important positions C_6 and $C_{3'}$ to substitute the new functional groups to generate new flavonoids with higher activity than luteolin isolating from leaf of Perilla ocymoides L. The $QSAR_{PCA-ANN}$ model with architecture I(6)-HL(9)-O(1) has the good applicability for flavonoids. The biological activities resulting from $QSAR_{PCA-ANN}$ model turn out to be in good agreement with those from experimental data. The QSAR models described in the present paper for diverse flavonoids may be useful for in vitro toxicity assessment. This work established the different models QSAR that may prove to be useful for guiding the rational search of new therapeutic agents for cancer diseases.

References

1. Do Tat Loi (2006) Medicinal Plants and Drugs from Vietnam. Publisher of Medicine, Vietnam.
2. Manjinder Singh, Maninder Kaur, Om Silakari (2014) Flavones: an important scaffold for medicinal chemistry. European Journal of Medicinal Chemistry 84: 206-239.
3. Debarshi Kar Mahapatra, Sanjay Kumar Bharti, Vivek Asati (2015) Anti-cancer chalcones: Structural and molecular target perspectives. European Journal of Medicinal Chemistry 98: 69-114.
4. Lovro Ziberna, Stefano Fornasaro, Jovana Čvorović, Federica Tramer, Sabina Passamonti (2014) Polyphenols in Human Health and Disease 1: 489-511.
5. R Vidya Priyadarsini, R Senthil Murugan, S Maitreyi, K Ramalingam, D Karunagaran, et al. (2010) The flavonoid quercetin induces cell cycle

arrest and mitochondria-mediated apoptosis in human cervical cancer (HeLa) cells through p53 induction and NF-κB inhibition. European Journal of Pharmacology 649(1-3): 84-91.

6. Bożena Pawlikowska Pawlęga, Halina Dziubińska, Elżbieta Król, Kazimierz Trębacz, Anna Jarosz Wilkołazka, et al. (2014) Interaction of a quercetin derivative - lensoside Aβ with liposomal membranes. Biochimica et Biophysica Acta (BBA) - Biomembranes, 1838(1): 254-265.
7. Nathan M Gavin, Michael J Durako (2012) J. Experimental Marine Biology and Ecology 32(40): 416-417.
8. Iris S L Lee, Mary C Boyce, Michael C Breadmore (2012) Extraction and on-line concentration of flavonoids in Brassica oleracea by capillary electrophoresis using large volume sample stacking. J. Food Chemistry 133(1): 205–211.
9. Bui Thi Phuong Thuy Pham Van Tat (2012) Vietnam Journal of Chemistry 5(50): 550-556.
10. Bui Thi Phuong Thuy Pham Van Tat (2012) Vietnam Journal of Chemistry 50(5A): 203-208.
11. Pham Van Tat (2009) Vietnamese Journal of Chemistry and Application 14: 43-46.
12. Si Yan Liao, Jin Can Chen, Li Qian, Yong Shen, Kang Cheng Zheng, et al. (2005) J Bioorganic and Medicinal Chemistry 13: 6045-6053.
13. Si Yan Liao, Jin Can Chen, Li Qian, Yong Shen, Kang Cheng Zheng, et al. (2008) QSAR, action mechanism and molecular design of flavone and isoflavone derivatives with cytotoxicity against HeLa. J European Journal of Medicinal Chemistry 43(10): 2159-2170.
14. I Li Chen, Chen JY, Shieh PC, Chen JJ, Lee CH, et al. (2008) Synthesis and antiproliferative evaluation of amide-containing flavone and isoflavone derivatives. J Bioorganic and Medicinal Chemistry 16(16): 7639-7645.
15. TC Wang, IL Chen, PJ Lu, CH Wong, CH Liao, et al. (2005) Synthesis, antiproliferative, and antiplatelet activities of oxime-and methyloxime-containing flavone and isoflavone derivatives, Bioorganic & Medicinal Chemistry. Bioorg Med Chem 13(21): 6045-6053.
16. Pham Van Tat (2009) Development of Quantitative Structure-Activity Relationships (QSARs) and Quantitative Structure-Property Relationships (QSPRs). Publisher of Natural Science and Technology, Hanoi.
17. XLSTAT version (2014) Copyright Addinsoft 1995-2014, USA.
18. Jung, SK, K McDonald (2011) Visual Gene Developer: a fully programmable bioinformatics software for synthetic gene optimization. BMC Bioinformatics 12(1): 340.
19. Wold S (1995) PLS for multivariate linear modelling. In: van de Waterbeemd H, QSAR: Chemometric Methods in Molecular Design. Wiley-VCH, Weinheim, Germany 2: 195-218.
20. OriginLab tutorials, OriginLab Corporation, Northampton, MA 01060, USA (2015).
21. HyperChem Release (2008) Hypercube Inc USA.
22. Bastien P, Esposito Vinzi V, Tenenhaus M (2005) PLS Generalised Regression. Computational Statistics and Data Analysis 48: 17-46.
23. Pham Van Tat (2017) Development of New Anticancer Agents From Leaf of Plants In Viet Nam, LAP Lambert Academic Publishing, OmmiScriptum GmbH & Co KG, Germany.
24. Bui Thi Phuong Thuy Pham Van Tat (2012) Vietnam Journal of Chemistry 50(5A): 203-208.
25. Bui Thi Phuong Thuy, Nguyen Thi Ai Nhung, Tran Duong, Phung Van Trung, Nguyen Minh Quang, et al. (2016) Prediction of anticancer activities of cynaroside and quercetin in leaf of plants *Cynara scolymus L* and *Artocarpus incisa L* using structure-activity relationship. Theoretical and Computational chemistry, Cogent Chemistry, Taylor & Francis, Cogent Chemistry 2(1).

DOI: 10.19080/OMCIJ.2017.04.555638