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*, Bui Thi Phuong Thuy1, Tran Duong2, Phung Van Trung4, Hoang Thi Kim Dung4 and Pham Nu Ngoc Han3

1Faculty of Chemistry, Hue University of Science, Asia
2Faculty of Chemistry, Hue University of Education, Asia
3Faculty of Science and Engineering, Hoa Sen University, Asia
4Institute 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 QSARPCA-ANN relationships. The best multiple linear model QSARMLR (with k = 6) values R²_train of 0.854 and R²_pred of 0.812, and QSARPCR (with k = 6) values R²_train of 0.937 and R²_pred of 0.889 were found by using the multiple linear regression technique. The artificial neural network QSARPCA-ANN with architectural style I (6)-HL (9)-O (1) represented the values R²_train of 0.993 and R²_pred of 0.971. In the case the incorporated model QSARPCA-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: QSARMLR, QSARPCR, and QSARPCA-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 silico method has become an important tool nowadays. In silico 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 silico model 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 silico 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/μM 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 silico 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 GI50/μM for Hela cell line (GI50 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 logGI50 is the subsequent dependent variable that defines the biological parameter for QSAR model.

\[ pGI_{50} = -\log GI_{50} \]  

(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 QARIS 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]:

\[ Y = \sum_{i} a_{x_{i}} + b \]  

(2)

Where Y is the experimental activity pGI_{50, exp}, x_{i} is k\textsuperscript{th} molecular descriptor.

Values R\textsuperscript{2}\textsubscript{train} and R\textsuperscript{2}\textsubscript{pred} are calculated by

\[ R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \]  

(3)

Here Yi, Yi are experimental pGI_{50, exp} and predicted pGI_{50, pred} value; \bar{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]:

\[ ARE,\% = 100 \left( \frac{pGI_{50,exp} - pGI_{50,pred}}{pGI_{50,exp}} \right) \]  

(4)

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

\[ MARE,\% = \frac{100}{N} \left( \frac{pGI_{50,exp} - pGI_{50,pred}}{pGI_{50,exp}} \right) \]  

(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 PCs 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

Figure 1: Molecular skeleton: a) flavone and b) isoflavone.
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 Q SARIS [16] using the optimized geometries. The molecular descriptors were used to construct the multiple linear regression (Q SAR MLR), principal component regression (Q SAR PCB-ANN) and artificial neural network (Q SAR PCANN) model [4,5].

b. Development of Q SAR MLR model: Before conducting the Q SAR MLR model, the activity values GI₅₀(µM) are transformed into the values pGI₅₀ to adapt the statistical properties. The values pGI₅₀ (µM) are most appropriate value for modelling the relationships between molecular descriptors and activities. The Q SAR MLR models were established by using the relationship of the geometric predictors and biological activities pGI₅₀ [16]. The Q SAR 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 Q SAR 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 GI₅₀ (µM) of flavones and isoflavones [2].

| Substance | Name | Substitutive site | Substitutes R | GI₅₀ (µM) |
|-----------|------|------------------|---------------|----------|
|           |      |                  |               |          |
| Training set for establishing Q SAR models |
| Fla1      | 1a -1 | flavone          | C₁            | -OCH₃CH₂-NOH | 2.0 |
| Fla2      | 1b    | flavone          | C₁            | -OCH₃CH₂-NOH | 1.2 |
| Fla5      | 4a    | flavone          | C₁            | -OCH₃CH₂-NOH | 2.0 |
| Fla6      | 5a    | flavone          | C₁            | -OCH₃CH₂-NOH | 0.9 |
| Fla7      | 6a    | flavone          | C₁            | -OCH₃CH₂-NOH | 2.2 |
| Isofla8   | 7a    | isoflavone       | C₁            | -OCH₃CH₂-NOH | 8.5 |
| Fla10     | 8a    | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 2.1 |
| Fla11     | 9a    | flavone          | C₁            | -OCH₃C(4-CH₃O-C₆H₄)=NOH | 2.0 |
| Fla12     | 3b    | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 0.8 |
| Fla13     | 10a   | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 1.6 |
| Fla14     | 11a   | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 1.0 |
| Fla15     | 4b    | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 2.0 |
| Fla16     | 5b    | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 2.0 |
| Fla17     | 12a   | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 2.0 |
| Isofla18  | 13a   | isoflavone       | C₁            | -OCH₃C(4-C₆H₄)=NOH | 9.0 |
| Isofla19  | 14a   | isoflavone       | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 7.8 |
| Isofla20  | 15a   | isoflavone       | C₁            | -OCH₃C(4-C₆H₄)=NOH | 7.6 |
| Fla21     | 16a   | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 1.6 |
| Fla23     | 18a   | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 2.0 |
| Fla24     | 19a   | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 2.4 |
| Fla25     | 20a   | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 2.3 |
| Fla26     | 21a   | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 2.0 |
| Fla27     | 22a   | flavone          | C₁            | -OCH₃C(4-C₆H₄)=NOH | 6.6 |
| Fla28     | 3a    | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 2.7 |
| Fla29     | 24a   | flavone          | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 2.5 |
| Isofla30  | 25a   | isoflavone       | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 8.2 |
| Isofla31  | 26a   | isoflavone       | C₁            | -OCH₃C(4-F-C₆H₄)=NOH | 6.4 |
| Test set for validating Q SAR models |
| Fla3      | 2a-3  | flavone          | C₁            | -OCH₃CH₂-NOH | 2.0 |
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 \( R^2_{\text{train}} \), \( R^2_{\text{pred}} \) 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 \( R^2_{\text{train}} \), SE and \( R^2_{\text{pred}} \). The values \( R^2_{\text{pred}} \) 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 2: The QSAR MLR models (k from 2 to 10) with change of values \( R^2_{\text{train}} \), \( R^2_{\text{pred}} \) and SE.

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

Table 3: Statistical values and valuable contribution percentages MP\( x_k \),% and GMP\( x_k \),% for 2D and 3D molecular descriptors in the models QSAR MLR (with \( k \) of 5, 6 and 7).

| Variable | \( x_k \) | QSAR MLR | MP\( x_k \),% | GMP\( x_k \),% |
|---|---|---|---|---|
| \( R^2_{\text{train}} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| 0.832 | 0.854 | 0.836 |
| 0.820 | 0.841 | 0.820 |
| 0.365 | 0.342 | 0.365 |
| 0.756 | 0.812 | 0.721 |
| 3.883 | 8.509 | 4.790 |
| \( R^2_{\text{adj}} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( SE \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( R^2_{\text{pred}} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| 0.022 | - | - | - | 27.945 | 18.636 | 19.005 |
| 0.820 | - | - | 0.0143 | - | - | 0.433 |
| 3.416 | 2.8540 | 3.588 | 24.043 | 25.862 | 25.908 | 25.271 |
| \( \text{ABSQ} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| - | - | - | 24.203 | 23.890 | 16.031 |
| \( \text{ABSQon} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{MaxQp} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{MaxNeg} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{SdO} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{ka2} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{LogP} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{Ovality} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{SdssC} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
| \( \text{ka3} \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) | \( k = 5 \) | \( k = 6 \) | \( k = 7 \) |
The QSAR_{MLR} models (with k of 2 to 10) that were arranged in an orderly change of R^{2}_{\text{train}} > SE and R^{2}_{\text{pred}} as given in (Table 2). The values of R^{2}_{\text{train}} and R^{2}_{\text{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}_{\text{train}} of 0.854 and R^{2}_{\text{pred}} 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_{\text{xp}}% and GMP_{\text{xp}}% [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_{\text{xp}}% and GMP_{\text{xp}}% of each independent variable for 3 models is determined by the formula

\[ MP_{\text{xp}}\%, \, \text{GMP}_{\text{xp}}\% = \frac{1}{N} \sum_{i=1}^{n} \left(100 \left| \frac{p_{i,x_i}}{C_{i,x_i}} \right| \right) \quad \text{with} \quad C_{i,x_i} = \sum_{i=1}^{n} p_{i,x_i} \]  

(6)

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

\[ GMP_{\text{xp}}\% = \frac{1}{n} \sum_{m=1}^{3} MP_{\text{xp}}\% \]  

(7)

With n number of models

The contribution percentages GMP_{\text{xp}}% in Table 3 depicted the important level of 2D and 3D molecular descriptors for flavonoids isolated from leaves of Perilla ocymoides L and Glucose max L. The best QSAR_{MLR} model with simulated annealing variable selection mode by using PCR model [17]. The 6 independent variables MaxQp, Sd0, 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):

\[ pGI_{\text{exp}} = 5.48356 + 0.38027 \times PC_1 - 0.11868 \times PC_2 + 0.34789 \times PC_3 + 0.06995 \times PC_4 + 0.21850 \times PC_5 + 0.35057 \times PC_6 \]  

(8)

With n = 26; R^{2}_{\text{train}} = 0.937; R^{2}_{\text{adj}} = 0.9106, R^{2}_{\text{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_{exp} and pGI_{pred} values is pointed out in (Figure 2).

**Figure 2:** The nature of principle component and correlation between pGI_{exp} and pGI_{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 Glucose max L. Organic & Medicinal Chem IJ. 2017; 4(3): 555638. DOI: 10.19080/OMC1.2017.04.555638
d. Building QSA[PCA-ANN] model: The QSA[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) PC1, PC2, PC3, PC4, PC5 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 pGI50; 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 learning rate of 0.7; the goal monitoring error MSE = 0.000816 with 10,000 iteration. After training the QSA[PCA-ANN] model with architecture I(6)-HL(9)-O(1) pointed out the values R2train of 0.993 and R2pred of 0.971. But in the case the QSA[PC] model gave values R2train = 0.937 and R2pred = 0.889; and the QSA[MLR] model (with k = 6) gave values R2train of 0.854 and R2pred 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-Aluofolien 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

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 H2SO4/EtOH or FeCl3/EtOH.

iv. Nuclear magnetic resonance spectrum (NMR) 1H-NMR
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 $R^{2}_{\text{pred}}$; 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).

| Substance | Ref. | pGI_{exp} | pGI_{pred} | ARE,% |
|-----------|------|-----------|------------|-------|
|           |      | QSAR_{MLR} | QSAR_{PCR} | QSAR_{PCA-ANN} | QSAR_{MLR} | QSAR_{PCR} | QSAR_{PCA-ANN} |
| Fla3      | [8,12,13] | 5.699 | 5.532 | 5.632 | 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 | 5.764 | 1.511 | 0.172 | 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

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}.

The luteolin structure was identified by using the different spectra such: $^1$H-NMR (DMSO-d$_6$, 500 MHz, 8 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$_5$); 7.4 (1H; s; H$_7$); 6.89 (1H; d; J = 8Hz, H$_8$); 7.41 (1H; d; J = 8Hz, H$_7$); 12.95 (1H, s; C5-OH); 9.4 (1H, s, C$_3$-OH); 9.9 (1H, s, C$_3$-OH); 10.84 (1H, s, C7-OH). The $^{13}$C-NMR spectrum was employed to have more information such as combining $^{13}$C-NMR (DMSO-d$_6$, 125 Hz) with spectrum DEPT, HSQC, HMBC: 6163.1 (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(C$_2$); 145.7(C$_3$); 149.7(C$_4$); 116.0(C$_5$); 118.9(C$_6$). 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$_7$-C$_3$-C$_5$-C$_7$.
The use of computational methods constructed successfully the in silico models with relationships between the 2D, 3D molecular descriptors and anti-cancer activities GI50 (µM) of flavonoids. The QSAR 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 silico model QSAR also found out helpfully the most important positions C$_{17}$, C$_{2}$ 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 model with architecture I(6)-HL(9)-O(1) has the good applicability for flavonoids. The biological activities resulting from QSAR 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. Debashri 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 Čerić, Federica Tramer, Sabina Passamonti (2014) Polyphenols in Human Health and Disease 1: 489-511.
5. R Vidy Priyadarshini, R Senthil Murugan, S Maitreyi, K Ramalingam, D Karunagaran, et al. (2010) The flavonoid quercetin induces cell cycle

**Table 5:** The anticancer activities pGI$_{50}$ of 5 new flavonoids resulting from QSAR-ANN model.

| New substance | C$_{17}$ | 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        | GR$_{6}$-CO- | H       | NO$_{2}$ | 6.155 | QSAR-PCA-ANN        |
| flav-3        | GR$_{6}$-CO- | NO$_{2}$ | H       | 6.158 | QSAR-PCA-ANN        |
| flav-4        | NO$_{2}$ | NO$_{2}$ | H       | 6.488 | QSAR-PCA-ANN        |
| flav-5        | NO$_{2}$ | NO$_{2}$ | NO$_{2}$ | 6.538 | QSAR-PCA-ANN        |

**Figure 6:** Comparison between values pGI$_{50}$ of five new flavonoids with luteolin.

**Conclusion**

The use of computational methods constructed successfully the in silico models with relationships between the 2D, 3D molecular descriptors and anti-cancer activities GI50 (µM) 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 QASR models can be accepted in uncertainty range of experimental measurements. Consequently, the models QSAR$_{MLR}$, QSAR$_{PLS}$ 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 is substituted at the vacant positions C$_{17}$, C$_{2}$ and C$_{3}$.
6. Bożena Pawlikowska Pawlega, Halina Dziubińska, Elżbieta Król, Kazimierz Trębacz, Anna Jarosz Wilkołazka, et al. (2014) Interaction of a quercetin derivative - lensoside A8 with liposomal membranes. Biochimica et Biophysica Acta (BBA) - Biomembranes, 1838(1): 254-265.

7. Nathan M Gavin, Michael J Duralo (2012) J. Experimental Marine Biology and Ecology 32(40): 416-417.

8. Iris SL 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. 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, 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 synthetice 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, Northampon, 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, OmniScriptum 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 Ngh, 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).