2D QSAR Study on Saponins of *Pulsatilla koreana* as an Anticancer Agent

Bhushan A Baviskar, Sharada L Deore, Sagar Alone
Government College of Pharmacy, Amravati-444 604, Maharashtra, INDIA.

ABSTRACT

In the present communication, total seventeen saponin molecules were collected and subjected to quantitative structure-activity relationship (QSAR) analyses. 2D QSAR model is developed by using multiple linear regression method against 4 different cell lines. QSAR model was generated by using training set of 11 and test set of 6 molecules having correlation coefficient ($r^2$), significant cross validated correlation coefficient ($q^2$) and F-test (For statistical significance) is as given below (A-549: $r^2$: 0.9281, $q^2$: 0.8691, F-test: 51.6079), (SK-OV-3: $r^2$: 0.9564, $q^2$: 0.9184, F-test: 85.7357), (SK-MEL-2: $r^2$: 0.9160, $q^2$: 0.8285, F-test: 43.6084), (HCT15: $r^2$: 0.9203, $q^2$: 0.8357, F-test: 46.1887). In this QSAR study Alignment independent descriptors such as $T_2\_C_7$, $T\_O\_O_5$ and physicochemical descriptors like Chain path count such as 6 chain count and Chi chain such as Chi 6 chain were most responsible descriptors for anticancer activity.

Key words: QSAR, *Pulsatilla koreana*, Cytotoxic activity, Multiple linear regressions.

Correspondence: Bhushan A Baviskar
Government College of Pharmacy, Amravati-444 604, Maharashtra, INDIA.
E-mail: baviskarbhushan@gmail.com; sharudeore_2@yahoo.com
DOI: 10.5530/pc.2019.1.2.

INTRODUCTION

Cancer is one of the dared disease characterized by progressive, abnormal and uncontrolled propagation of tissues. Currently, cancer is most leading cause of death of human being in world.\(^{1,2}\) A variety of anticancer agents have been developed over the last decades. Still, there is no doubt that further research into the design of new anticancer compounds with low toxicity and higher selectivity is needed.\(^{3,4}\) Natural products or herb extracts have been emerged as a novel approach to control a variety of diseases, including cancers. They are the basis for the traditional medicine philosophies and practices in China, India and tribal peoples.\(^6\) The *Pulsatilla koreana* (PKE) belongs to the family Ranunculaceae, in Korea is an important herbal medicine that has been used to treat amoebic dysentery, malaria and also demonstrate antifungal, antibiotic properties, anti-inflammatory, anti-acute activities.\(^{6,9}\) In addition, several studies have been reported that the saponins, mainly the oleanane and lupane-type saponins in PKE have anticancer effects in human melanoma, colon, hepatocellular and lung carcinoma.\(^{11-13}\)

Presently, computational methods and molecular modeling are the inseparable parts in drug design and discovery and no one can talk about drug design without having a bit comprehension about computational chemistry.\(^{14-18}\) Amongst the various computational technique exist in drug design, Quantitative structure-activity relationships (QSARs) have been found the major popularity.\(^9\) QSAR, as one of the most significant application in chemometrics, gives information which is helpful for design of new molecule acts on specific target.\(^{19-22}\) QSAR have been developed since 1963, when Hansch and coworker developed quantitative relationships between activity and the octanol-water partition coefficient.\(^{23}\) It is assumed that the sum of steric effects on the steric, electronic and hydrophobic interaction of the compounds with their receptor determines the biological activity. Another criterion in constructing the QSAR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number.\(^{24}\) These include both 2D (two dimensional) and 3D (three-dimensional) QSAR methods. The major differences of these methods can be analyzed from two viewpoints: (1) the structural parameters that are used to characterize molecular identities and (2) the mathematical procedure that is employed to obtain the quantitative relationship between a biological activity and the structural parameters.

The present communication deals with 2D QSAR analysis of 17 saponin molecules with natural resource in which eight lupane-type and nine oleanane-type by using multiple linear regression method against 4 different cell lines for anticancer activity.

MATERIALS AND METHODS

2D QSAR Methodology

Molecular structure generation and Data set:

A dataset of 17 saponins molecules isolated from the *Pulsatilla koreana* were taken from the literature and used for QSAR analysis.\(^{11}\) The 2D-QSAR models were generated for this series using multiple linear regression (MLR) methods against four human solid tumor cell lines (A-549, SK-OV-3, SK-MEL-2 and HCT-15), those models which come out with promising results are discussed here. The structures and their inhibitory activities (ED\(_{50}\) values) are listed in Table 1. All the molecular modeling and statistical analysis were carried out by using VLife MDS 4.0 software. The structures of the compounds were built using molecular sketching facilities offer in the modeling environment of VLife engine on ASUS laptop with a Dual core processor and Windows 7 operating system. Energy minimization and batch optimization was performed using Merck Molecular force field. All the molecules were initially optimized and then used for the calculation of descriptors and further QSAR study.\(^{25-26}\)

Selection of training and test set

The dataset of 17 molecules was divided into the training set (11 compounds) and test (6 compounds) for generating multiple linear regressions (MLR) 2D QSAR models.\(^{27}\) Therefore, the care was taken in such a way that biological activities of all compounds in test lie within the maximum and minimum value range of biological activities of training set of compounds. The Uni-Column Statistics of test and training sets further reflected the correct selection of test and training sets. A Uni-Column statistics for training set and test set were generated to check correctness of selection criteria for trainings and test set molecules Table 1.

The maximum and minimum value in training and test set were compared in a way that:
2D-QSAR study requires the calculation of molecular descriptors. A large number of theoretical 2D descriptors such as individual, Mol. Wt., Volume, XlogP, smr; retention index (chi), atomic valence connective index (chiv), path count, chi chain, chiv chain, chain path count, cluster, path cluster, kappa, physiochemical such as Estate Numbers, Estate contributions, Polar Surface Area, Element Count, Dipole moment, Hydrophobicity XlogP, Hydrophobicity SlogP; distanced based topological, alignment-independent and atom-type have been computed for these geometrically optimized structures from the chemical structures of the compounds referred to above with a view to develop structure–activity.25

Regression analysis

The 2D-QSAR models were generated with stepwise multiple linear regression (MLR) analyses using forward–backward technique set as 0.0 kcal/mol cut-off, applying auto scaling. It relates the dependent variable y (biological activity) to a number of independent variables xᵢ (molecular descriptor) by using linear equations. MLR is the traditional

1. The maximum activity of test set should be less than or equal to the maximum value of training set.

2. The minimum value of test set should be higher than or equal to the minimum value of training set.

This observation showed that test set was interpolative and derived within the minimum–maximum range of training set. The mean and standard deviation of ED₅₀ values of sets of training and test provide insights to the relative difference of mean and point density distribution (Along mean) of the two sets (www.Vlifesciences.com).26

QSAR models were generated by a training set of 11 molecules for each model. And a test set of 6 molecules with uniformly distributed biological activities. The structures of all the compounds referred to above with a view to develop structure–activity.

Descriptor calculation

The basis of energy minimization is that the drug binds to effectors/receptors in the most stable form, i.e., the minimum energy form. For 2D-QSAR study requires the calculation of molecular descriptors. A large number of theoretical 2D descriptors such as individual, Mol. Wt., Volume, XlogP, smr; retention index (chi), atomic valence connective index (chiv), path count, chi chain, chiv chain, chain path count, cluster, path cluster, kappa, physiochemical such as Estate Numbers, Estate contributions, Polar Surface Area, Element Count, Dipole moment, Hydrophobicity XlogP, Hydrophobicity SlogP; distanced based topological, alignment-independent and atom-type have been computed for these geometrically optimized structures from the chemical structures of the compounds referred to above with a view to develop structure–activity.25

Regression analysis

The 2D-QSAR models were generated with stepwise multiple linear regression (MLR) analyses using forward–backward technique set as 0.0 kcal/mol cut-off, applying auto scaling. It relates the dependent variable y (biological activity) to a number of independent variables xᵢ (molecular descriptor) by using linear equations. MLR is the traditional
Validation of QSAR model

The generated QSAR model was validated for predictive ability inside the model by using cross validation (leave-one-out—LOO) for q² and external validation, which is more robust alternative method by dividing the data into training set and test set.² For calculating q², each molecule in the training set was eliminated once and the activity of the eliminated molecule was predicted by using the model developed by the remaining molecules. The q² was calculated using the equation Eq. 2 which describes the internal stability of a model.

\[ q^2 = 1 - \frac{\sum (y_i - y_i^{\text{pred}})^2}{\sum (y_i - y_{\text{mean}})^2} \]

Where \( y_i \) and \( y_i^{\text{pred}} \) are the actual and predicted activity of the \( i \) th molecule in the training set, respectively and \( y_{\text{mean}} \) is the average activity of all molecules in the training set.

The predictive ability of the selected model was also confirmed of test set compounds which is also denoted with \( \text{pred}_{r^2} \). The \( \text{pred}_{r^2} \) value is calculated as follows

\[ \text{pred}_{r^2} = 1 - \frac{\sum (y_i - y_i^{\text{pred}})^2}{\sum (y_i - y_{\text{mean}})^2} \]

Where \( y_i \) and \( y_i^{\text{pred}} \) are the actual and predicted activity of the \( I \) th molecule in the training set, respectively and \( y_{\text{mean}} \) is the average activity of all molecules in the training set. The statistical significance of selected 2D-QSAR model was further supported by the 'fitness plot' obtained, this is a plot of observed versus predicted activity of training and test set compounds and provides an idea about how well the model was trained and how well it predicts the activity of the external test set (Figure 1). The contribution chart for the significant model is presented in Figure 2, which gives the percentage contribution of the descriptors used in deriving the model.

RESULTS AND DISCUSSION

The aim of our study was to developed 2D-QSAR model for a series of analogs of Pulsatilla koreana saponin by using MLR method. 2D-QSAR equations were selected by optimizing the statistical results generated along with variation of the descriptors in this model. The fitness/pattern plots were also generated for evaluating the dependence of the biological activity on various different types of the descriptors. The frequency of use of a particular descriptor in the population of equations indicated the relevant contributions of the descriptors.

The best regression equation obtained is represented in Equation:

- For (A-549): \( E_{\text{DA}} = + 55.3309(\pm 6.5096) + 23.9224(\pm 8.9012) \times 6\text{ChainCount} - 616.7280 \)
- For (SK-OV-3): \( E_{\text{DA}} = + 53.2371(\pm 0.5713) + 553.1007(\pm 130.2520) \times 6\text{ChainCount} - 612.2515 \)
- For (SK-MEL-2): \( E_{\text{DA}} = + 53.9707(\pm 7.2274) + 255.87(\pm 2.5587) \times 6\text{ChainCount} - 593.8088 \)
- For (HCT15): \( E_{\text{DA}} = + 53.7128(\pm 6.9414) + 3583(\pm 4.2574) \times 6\text{ChainCount} - 484.2732 \)

The results of statistical parameters are summarized in Table 3. From the QSAR studies it was found that the Alignment Independent category in which \( T_\text{2-C}_7 \) descriptor is most powerful descriptor responsible for anticancer activity to all 4 types of cell lines at range of 70% - 80% A-549, SK-OV-3, SK-MEL-2 and HCT15. This descriptor signifies No. of double bounded atoms (i.e.- Any double bonded atoms, T-2) separated from carbon atom by 7 bonds in a molecule. \( T_\text{O-O}_5 \) descriptor is responsible for anticancer activity in SK-MEL-2 and HCT15 cancer cell line at range 20%-30% and signifies No. of oxygen atoms (Single double
Table 3: Statistical results of 2D QSAR equation generated by MLR method.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>A-549</th>
<th>SK-OV-3</th>
<th>SK-MEL-2</th>
<th>HCT-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>DOF</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>r²</td>
<td>0.9281</td>
<td>0.9554</td>
<td>0.9160</td>
<td>0.9203</td>
</tr>
<tr>
<td>q²</td>
<td>0.8691</td>
<td>0.9187</td>
<td>0.8285</td>
<td>0.8357</td>
</tr>
<tr>
<td>F test</td>
<td>51.6079</td>
<td>85.7357</td>
<td>43.6084</td>
<td>46.1887</td>
</tr>
<tr>
<td>r² se</td>
<td>34.4579</td>
<td>26.6109</td>
<td>37.7639</td>
<td>36.2697</td>
</tr>
<tr>
<td>q² se</td>
<td>46.4909</td>
<td>35.9315</td>
<td>53.9485</td>
<td>52.0702</td>
</tr>
<tr>
<td>pred_r²</td>
<td>0.9538</td>
<td>0.9686</td>
<td>0.8950</td>
<td>0.8899</td>
</tr>
<tr>
<td>pred_r² se</td>
<td>31.0692</td>
<td>25.8051</td>
<td>47.7110</td>
<td>48.3659</td>
</tr>
</tbody>
</table>

N - Number of molecules, K - Number of descriptors in a model, DOF - Degree of freedom (higher is better), r² - Coefficient of determination (> 0.7), q² - Cross-validated r (>0.5), pred_r² - r for external test set (>0.5), F-test - F-test for statistical significance of the model (higher is better, for same set of descriptors and
descriptor values).

Figure 1: Fitness plot of model A-549 cell line, SK-OV-3 cell line, SK-MEL-2 cell line and HCT-15 cell line.

Figure 2: Contribution plot of model for different cancer cell line A-549, SK-OV-3, SK-Mel-2 and HCT-15.

CONCLUSION

In order to determine antitumor activities of various Pulsatilla koreana saponins may be treated statistically to find out the structural characteristics which are essential for high activity. 2D-QSAR has been carried out on Pulsatilla koreana saponins analogue. It was found that all descriptors have positive correlation with the activity. Increase in all descriptor will help to increase the antitumor activity of Pulsatilla koreana saponins analogue. Existence of free carboxylic group is at C-28 position is most responsible for cytotoxic activity and hydroxyl group at C-23 had a negative effect on the cytotoxic activity. It is due to electron donating effect of two oxygen atom towards C-3 of aglycon. The results obtained from this 2D-QSAR study are in agreement with the observed SAR of Pulsatilla koreana saponin studied. There is difference in cytotoxic between oleanean and lupane saponin generally cytotoxicity of lupane type saponin were much weaker than those of oleanine type saponin that means oleanan type Pulsatilla koreana saponin analogues having more antitumor activity than that of lupane type. Statistical significance of these models was further supported by ‘fitness plot’ (Figure 1) obtained for each model, this is a plot of experimental versus predicted activity of training and test set compounds and provides an idea about how fit the model was trained and how well it predicts activity of external test set. Contribution charts for all the significant models are presented in Figure 2, which gives percentage contribution of descriptors used in deriving the QSAR models.

ACKNOWLEDGMENT

Authors are thankful to entire team of Vlife MDS for their constant support.

CONFICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

9. Campbell WE, Cragg GML, Powrie AH. Anemonin, protoanemonin and ranunculin...


