Pharm Methods, 2019; 10(1): 15-20.
A multifaceted peer reviewed journal in the field of Pharm Analysis and Pharmaceutics
www.phmethods.net | www.journalonweb.com/phm

Original Article

QSAR Studies of Triterpenoid Saponin Analogues for Nematicidal Activity

Bhushan A Baviskar*, Sharada L Deore, Priyanka Rathi

Government College of Pharmacy, Amravati, Maharashtra, INDIA.

ABSTRACT

2-D QSAR of triterpenoid saponin analogues with nematicidal activity performed by using three methods: Multiple Linear Regression (MLR), Partial Least Square (PLS) and Principle Component Regression (PCR). The overall degree of prediction of descriptor was found to be around 100% in all three models: MLR, PLS PCR. But, result of Multiple Linear Regression (MLR) analysis showed significant predictive power and reliability as compared to other two methods. The correlation coefficient r^2-0.8684 indicates 86.84% correlation between activities and molecular descriptors of training set compound. Cross validated regression coefficient q^2-0.82071 meaning that the prediction accuracy of QSAR is 82.07%. slogP descriptor having 100% positive correlation with the activity. This descriptor signifies log of the octanol/water partition coefficient (including implicit hydrogens). This property is an atomic contribution model that calculates logP from the given structure; i.e., the correct protonation state. Carboxyl group at position C-28 of aglycone is most responsible for nematicidal activity.

Keywords: Triterpenoid Saponins, Pulsatilla koreana roots, Nematicidal activity, QSAR, Descriptors, Multiple Linear regression (MLR), Partial least square (PLS), Principle component regression (PCR).

Correspondence
Mr. Bhushan A Baviskar
Government College of Pharmacy, Amravati-444 604, Maharashtra, INDIA.
E-mail: baviskarbhushan@gmail.com
DOI : 10.5530/phm.2019.1.3

INTRODUCTION

Saponins are the class of chemical compound this secondary metabolite found in natural sources particularly in plant species.[1] The presence of saponin has been reported in more than 100 families of plant and in few marine sources; Such as star fish and sea cucumber.[2] Saponins found primarily in dicotyledonous plant and also in monocots. The saponins have different values processed as drug and medicines, foaming agent, sweeteners, taste modifiers and cosmetics. They also have the different biological activity Antifungal, Antibacterial, Anti-inflammatory, Anticancer, nematicidal, Antiprotozoal, Chemo protective, immunomodulation, hypoglycaemic, Anti hepatotoxic, Antiphlogistic.[3]

Saponins are amphipathic glycosides containing one or more sugar chains on a triterpene and steroid aglycone backbone also called a sapogenin. The non-sugar or the aglycone unit of the saponin molecule is called the sapogenin or just the genin. The saponins can be divided into three major classes according to the structure of genin: Triterpene glycosides, steroid glycosides and steroid alkaloid glycosides are shown in the Figure 1.[3]

Biological activity of triterpenesaponins

There are many biological activities associated with saponins. Most of those arise from the chemical nature of saponins which have two constitutional moieties: The hydrophilic sugar and the usually lipophilic sapogenin. These properties are responsible for the interaction between saponins and cell membranes.[4]

Nematodes

The nematodes /or roundworms constitute the phylum Nematoda. They are a diverse an animal phylum inhabiting a very broad range of environments. Nematodes have successfully adapted to nearly every ecosystem from marine to fresh water, to soils and from the polarregions to the tropics, as well as the highest to the lowest of elevations. They are ubiquitous in freshwater, marine and terrestrial environments.

Plant-parasitic nematodes include several groups causing severe crop losses. The most common genera are

- Aphelenchoides (foliar nematodes),
- Ditylenchus, Globodera (potato cyst nematodes),
- Heterodera (soybean cyst nematodes),
- Longidorus, Meloidogyne (rootknot nematodes),
- Nacobbus,Pratylenchus (lesion nematodes),
- Trichodorus and Xiphinema (dagger nematodes).[5]

In the past, synthetic compounds have mainly been used for plant protection, but many of these pesticides have side effects including residues in plants, contamination of groundwater, the potential for adverse ecological impacts from pesticide use and the creation of a continuing need for the development of new nematode control strategies and products.[6] Recently, one alternative used has been to screen naturally occurring plant secondary compounds for appropriate nematicidal activity. Various nematicidal substances of plant origin such as triglycerides, sesquiterpenes, alkaloids, steroids, diterpenes and flavonoids, have been identified in this way.[7] These compounds can be developed for use as nematicides themselves or can serve as model compounds for the development of chemically synthesized derivatives with enhanced activity and reduced environmental impacts.[8]

In an effort for search of new nematicidal agents, with enhanced activity and reduced environmental impacts the present study focuses on 2D QSAR study on 21 Triterpenoidsaponin derivatives for quantifying the necessary structural and physicochemical requirements of this series of analogues having nematicidal activity by using Modern computational technology (QSAR) as potent nematicidal agent.

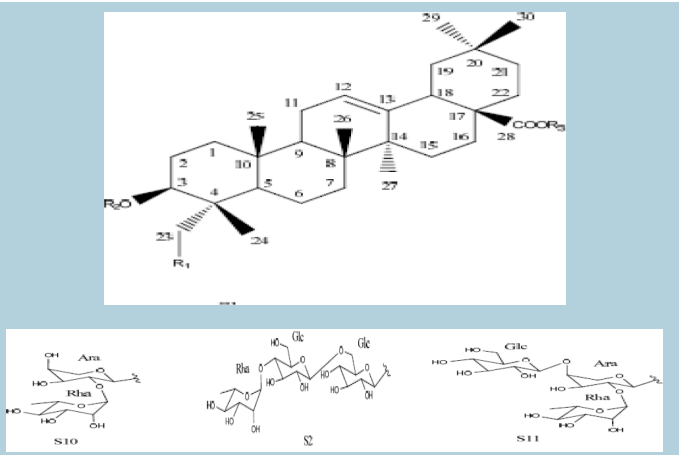
MATERIALS AND METHODS

Data set

The nematicidal activity data against the root-knot nematode (Meloidogyne incognita) were taken from reported work.[5] Since some compounds showed insignificant activity, such compounds were excluded from the dataset. Total 8 compounds (2, 5, 9, 12, 14, 20, 21 standard drug i.e fosthiazate) as shown in Figure 1 that are having significant activity are selected for the generic prediction of activity of thirteen triterpenoid saponin compounds (1, 3, 5, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17) as shown in Figure 2 isolated from same Pulsatilla koreana roots.[9] After generic prediction of activity of thirteen triterpenoid saponins, total 21 different analogues of triterpenoidal saponins selected to perform their 2D QSAR analysis.

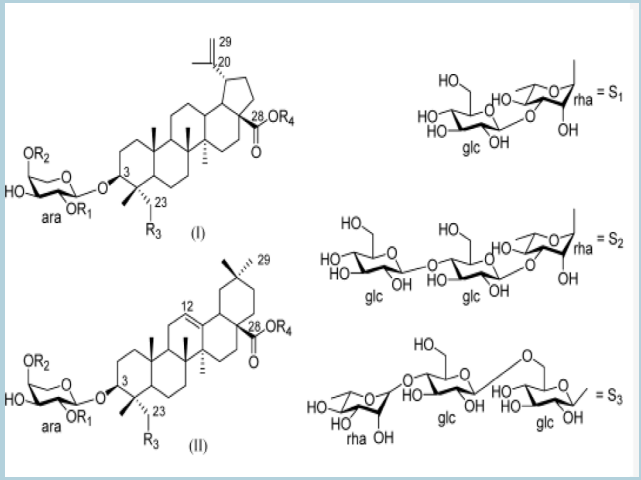
2D Structure of all above 21 triterpenoid saponins analogues were built using molecular sketching facilities offer in the modeling environment of Vlife engine on Lenovo Computer with a Intel-I3 core processor and Windows 7 operating system and then all 2D structures are converted to 3D by using VLife MDS (Molecular Design Suite) TM 3.5 software supplied by VLife Sciences Technologies Pvt. Ltd., Pune, India. And further subjected to energy minimization and optimization using batch energy minimization method, were conducted by using Merck Molecular Force Field. To get more minimization input of 100000 cycles given with the converse criteria (RMS gradient) of 0.01. The distance dependent function was kept at 1.0 with analytical gradient type.

The nematicidal activity data was given in LC_{50} (µg/mL) in the reported work.[5] Then this LC_{50} (µg/mL) values are converted to negative logarithm



Sr no	Saponins	R^1	R^2	R^3	Activities (PLC_{50})
1	2	H	S11	S2	4.023
2	5	H	S10	S2	4.051
3	9	H	S11	H	4.120
4	12	OH	S11	S2	3.870
5	14	OH	S1O	S2	3.894
6	20	OH	S10	H	4.154
7	21	OH	S11	H	4.097
8	Std. Drug Fosthiazate				4.138

Figure 1: Seven Triterpenoid Saponin analogues of Pulsatilla koreana root and fosthiazate as standard drug along with their activities in QSAR study



Sr.no.	Saponin	Aglycogen	R1	R2	R3	R4	Generic Prediction activities (PLC_{50})
1	1S	(I)	Rha	Glc	OH	S3	3.8995
2	3S	(I)	Rha	H	OH	S3	3.9655
3	5S	(I)	Rha	Glc	OH	H	4.0835
4	7S	(I)	S1	H	OH	H	4.0835
5	8S	(II)	S1	H	OH	H	4.0878
6	9S	(I)	Rha	Glc	H	H	4.1146
7	10S	(II)	Rha	Glc	H	H	4.1190
8	11S	(I)	Rha	H	OH	H	4.1495
9	13S	(I)	S1	H	H	H	4.1146
10	14S	(II)	S1	H	H	H	4.1190
11	15S	(I)	H	Glc	OH	H	4.1183
12	16S	(II)	H	Glc	OH	H	4.1227
13	17S	(II)	S2	H	OH	H	4.0218

Figure 2: Thirteen Saponin analogues of Pulsatilla koreana root and their generic prediction activities.

LC_{50} (µg/ml)= [-log(LC_{50}) × 10-6] was used as dependent variable for 2D and QSAR analysis.

Descriptor calculation

After the energy minimization and optimization of set of molecule, various 2D physicochemical descriptors (Total 239 descriptors) were calculated. Alignment independent descriptors (More than 700 descriptors): like molecular weight, log P, molar refractivity, retention index (chi), atomic valence connectivity index (chiv), path count, chain path count, cluster, path cluster, topological index, element counts, various Baumann alignment independent topological descriptor were also calculated using V-life MDS software. QSAR analysis were performed after removal of all the invariables coloumns, as they do not contributes to the QSAR.

Selection of training and test set

The dataset of 21 molecules was divided into training set (16 compounds) and test set (5 compounds) by Sphere Exclusion method for Multiple Linear Regression (MLR) with dissimilarity value 0.5, Mannual method for Partial Least Square (PLS) and Random method for Principal

Component Regression (PCA) using pLC_{50} activity field as dependent variable and various 2D descriptor as independent variable. The Uni-Column Statistics of test and training sets further reflected the correct selection of test and training sets, as the maximum of the training set was more than that of the test set and the minimum of the training set was less than or equal to that of the test set.

Regression Analysis

Dataset of 21 molecules was subjected to regression analysis using MLR, PCR, PLS as a model building methods. QSAR models were generated using pLC_{50} values as the dependent variable and various descriptor values as the independent variables. The cross correlation limit was set at 1 for MLR, 0.5 for PCR,1 for PCR term selection criteria as r^2 ,number of variables in final equation is at 3 for all methods, F-test 'in' at 4 and 'out' at 3.99, r^2 and F-test. Variance cut off was set at 0, scaling to auto scaling and number of random iterations to 10. Statistical measures were used for the evaluation of QSAR model were the number of compound in regression n, regression coefficient r^2, number of descriptors in a model k, F-test (Fisher test value) for statistical significance F, cross validated correlation coefficient q^2, predictive squared correlation coefficients pred_r^2, coefficient of correlation of predicted data set pred_r^2 se and standard error (SE) of estimation r^2 se and q^2 se.

2D QSAR analysis

Multiple Linear Regression (MLR) analysis

For the multivariate data analysis we use the standard method called multiple regression. It is also called as Ordinary least square regression (OLS). In this method by applying least squares curve fitting methods determine the value of regression coefficients. The data set having five times as many data points (molecule) as independent variable (Descriptors) is required for accurate results.

The regression equation: **Y= b1*x1+b2*x2+b3*x3+C**

Where, Y is dependent variable, the 'b' are regression coefficients for corresponding x is independent variable, C is a regression constant for intercept.[10]

In the present study QSAR model was developed using multiple regression by forward-backward variable selection method with pLC_{50} activity field as dependent variable and physiochemical descriptor as independent variable having cross correlation limit of 1 Selection of training set and test set was done by Sphere selection method.

b) Partial Least Square Regression method (PLS)

An extension of the multiple linear regression models called as partial least square regression. This model specifies linear relationship between a dependent variable Y and a set of predicted variables X's in its simplest form.

The regression equation: Y=b0+b1x1+b2x2+…….bpxp

b0 = regression coefficient for the intercept and bi values= regressΣion coefficient (for variable 1 through p) computed from data.

In PLS factors extracted from Y'XX'Y matrix represents prediction function. This typically extracted no. of such prediction functions will exceed the maximum no of Y and X variable. Furthermore, PLSR can be used as an exploratory analysis tool to select suitable predictor variable and to identify outliers before classical linear regression.[10]

In the present study QSAR model was developed using Partial Least Square Regression method by forward variable selection method with pLC_{50} activity field as dependent variable and physiochemical descriptor as independent variable having cross correlation limit of 1 Selection of training set and test set was done by manual selection method.

Principle Component Regression (PCR) Analysis

Principal components analysis provides a method for finding structure in such data sets. Put simply, it rotates the data into a new set of axes such that the first few axes reflect most of the variations within the data. By plotting the data on these axes, we can spot major underlying structures automatically. The value of each point, when rotated to a given axis, is called the principal component value. Principal Components Analysis selects a new set of axes for the data. These are selected in decreasing order of variance within the data. They are also perpendicular to each other. Hence the principal components are uncorrelated. Some components may be constant, but these will be among the last selected. The problem noted with MLR was that correlated variables cause instability. This process gives the modelling method known as Principal Components Regression. Rather than forming a single model, as with MLR, a model can be formed using 1, 2, .. components and a decision can be made as to how many components are optimal. If the original variables contained collinearity, then some of the components will contribute only noise. So long as these are dropped, the model can be we can guarantee that our models will be stable.[10]

In the present study QSAR model was developed using Principal component Regression method by forward variable selection method with pLC_{50} activity field as dependent variable and physiochemical descriptor as independent variable having cross correlation limit of 0.5 Selection of training set and test set was done by random selection method.

Validation of QSAR Model

Validation of QSAR study is important to test the internal stability and predictive ability of the QSAR models and was validated by the following procedure as given below. There are two types of validation.

Internal validation, External validation, Internal validation:

It was carried out using leave-one-out (q^2, LOO) method. For calculating q^2, 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 q^2 was calculated using the equation which describes the internal stability of a model.

$$\Sigma\,[yi(Act) - yi(Pred)]^2$$

$$q^2 = 1$$

$$\Sigma\,[yi - yi(mean)]^2$$

Where *yi (Act)* and *yi (Pred) are* the actual and predicted activity of the *i*th molecule in the training set, respectively and y mean is the average activity of all molecules in the training set.

External validation

The predictive ability of the selected model was also confirmed by external validation of test set compounds which is also denoted with pred_r^2. The pred_r^2 value is calculated as follows.

$$\Sigma\,[yi(Act) - yi(Pred)]^2$$

$$pred\_r^2 = 1 - \frac{\Sigma\,[yi(Act) - yi(Pred)]^2}{\Sigma\,[yi - yi(mean)]^2}$$

$$pred\_r2 = 1$$

$$\Sigma [yi - yi(mean)]^2$$

Where *yi*and *yi*are the actual and predicted activity of the *I* th molecule in the training set, respectively and y mean is the average activity of all molecules in the training set.[11]

RESULT AND DISCUSSION

The dataset of 21 molecules was divided into training set (16 compounds) and test set (5 compounds). Selection of molecules in the training set and test is a key and important feature of any QSAR model. 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. Generic predicted activity of thirteen saponins, experimental activity of eight saponins and predicted activity of twenty one saponins is shown in Table 1. 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 2).

In above QSAR models, r^2 is a correlation coefficient that has been multiplied by 100 gives explained variance in biological activity. Predictive ability of generated QSAR models was evaluated by q^2 employing LOO method. F value reflects ratio of variance explained by models and variance due to error in regression. High F value indicates that model is statistically significant. Low SE of estimation indicted by r^2 se and q^2 se suggested that models are statistically significant. Predictive ability of QSAR model was also confirmed by external validation of test set compounds denoted by pred_r^2 and it was found in agreement with accepted criteria of more than 0.3. Among these three models, MLR has come out with very good results as compare to other two models. Results of MLR analysis showed very good predictive ability as indicted by r^2, q^2, F-test, and pred_r^2 values (Table 3).

MLR equation come out as

pLC_{50} = + 0.0292 (± 0.0030) slog+ 4.0463 where Correlation Coefficient (r^2) is 0.8684, Cross validated Correlation Coefficient (q^2) is 0.8207, F test is 92.47 and degree of freedom of 14.

PLS equation come out as

pLC_{50} = + 0.0291 slogp+ 4.0462. where Correlation Coefficient (r^2) is 0.8593, Cross validated Correlation Coefficient (q^2) is 0.8081, F test is 85.5067 and degree of freedom of 14.

PCR equation come out as

pLC_{50} = + 0.0291 slogP+ 4.0466. In model 3rd Correlation Coefficient (r^2) is 0.8593, Cross validated Correlation Coefficient (q^2) is 0.7960, F test is 83.4500and degree of freedom of 14.

Figure 3, 5, 7 shows Contribution plot reveals that the individual descriptors like slog contributing positively as 100% to biological activity

Table 1: Generic predicted activity of thirteen saponins, experimental activity of eight saponins and predicted activity of twenty one saponins.

Sr No	Saponin	PLC_{50}			
		MLR, PLS, PCR Generic prediction activity upto 13 compound. and Experimental Activity 14...21 of 8 compound (PLC_{50})	MLR Predicted activity, for 21 compound.	PLS Predicted activity, for 21 compound.	PCR Predicted activity for 21 compound.
1	PK1S	3.8995	3.9044	3.9046	3.9049
2	PK3S	3.9655	3.9678	3.9679	3.9683
3	PK5S	4.0835	4.0813	4.0811	4.0815
4	PK7S	4.0835	4.0813	4.0811	4.0815
5	PK8S	4.0878	4.0855	4.0853	4.0857
6	PK9S	4.1146	4.1113	4.1110	4.1114
7	PK10S	4.1190	4.1155	4.1152	4.1156
8	PK11S	4.1495	4.1447	4.1444	4.1448
9	PK13S	4.1146	4.1113	4.1110	4.1114
10	PK14S	4.1190	4.1155	4.1152	4.1156
11	PK15S	4.1183	4.1148	4.1145	4.1149
12	PK16S	4.1227	4.1190	4.1187	4.1191
13	PK17S	4.0218	4.0220	4.0220	4.0223
14	Fosthiaz-ate	4.138	4.1580	4.1576	4.1580
15	PK2	4.023	3.9386	3.9387	3.9390
16	PK5	4.051	4.0020	4.0020	4.1490
17	PK9	4.120	4.1155	4.1152	4.1156
18	PK12	3.870	3.9086	3.9088	3.9091
19	PK14	3.894	3.9720	3.9721	3.9725
20	PK20	4.154	4.1489	4.1486	4.1490
21	PK21	4.097	4.0855	4.0853	4.0857

by models MLR, PLS and PCR respectively. Figure 4, 6, 8 shows Graph of Actual vs. predicted activities for training and test molecules by models MLR, PLS and PCR respectively.

As the cross-validated correlation coefficient (q^2) is used as a measure of reliability of prediction, the correlation coefficient suggests that our model is reliable and accurate. MLR model displays good predictivity in regular crossvalidation Table 4. Validation of QSAR Models revealed Slog *P* values for Model (MLR), Model 2(PLS) and Model 3(PCR) are + 0.0292 (± 0.0030), + 0.0291 and 0.0291 respectively. slogP descriptor signifies log of the octanol/water partition coefficient (including implicit hydrogens). This property is an atomic contribution model that calculates logP from the given structure; i.e., the correct protonation state.

CONCLUSION

It is concluded that statistically significant 2D-QSAR models were generated to predict structural features responsible for the nematicidal activities of triterpenoid saponin analogues against nematode- Melaidogyne incognita. Present study reveals that slogP as major contributing descrip-

Table 2: Uni-column statistics of the training and test sets for 2D-QSAR models.

MODELS	Training/ Test set	Average	Max	Min	Std. Dev.	Sum
MODEL 1 (MLR)	Training set	4.0497	4.1500	3.8700	0.0978	64.7952
	Test set	4.1050	4.1200	4.0834	0.0178	20.5251
MODEL 2 (PLS)	Training set	4.0457	4.1500	3.8700	0.0925	64.7311
	Test set	4.1177	4.1495	4.0835	0.0235	20.5887
MODEL 3 (PCR)	Training set	4.0552	4.1500	3.8700	0.0916	64.8839
	Test set	4.0872	4.1495	3.9655	0.0720	20.4359

Table 3: Statistics of the Models for MLR, PLS, PCR methods

Statistics	MLR	PLS	PCR
N	16	16	16
DF	14	14	14
r^2	0.8684	0.8593	0.8563
q^2	0.8207	0.8081	0.7960
F test	92.3742	85.5067	83.4500
r^2 se	0.0359	0.0359	0.0359
q^2 se	0.0419	0.0419	0.0428
pred_r^2	0.9967	0.9973	0.9978
pred_r^2se	0.0037	0.0043	0.0038
best_ran_r^2	0.06169	0.08385	0.04676
4best_ran_q^2	-0.25445	-0.19313	-0.30698
Best_ran_pred_r^2	-0.56592	-0.68037	0.39621

Where, N- Number of molecules, **K**- Number of descriptors in a model, **DF**- Degree of freedom (higher is better), r^2- Coefficient of determination (> 0.7), q^2- Cross-validated r (>0.5), **pred_r^2**- 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 compounds), **r^2_se, q^2_se, pred_r^2_se** = Error for r^2, q^2, pred_r^2 respectively.

Table 4: Internal and External validation data.

Sr. No	Validation	Model 1	Model 2	Model 3
1	Internal (q^2)	0.8207	0.8081	0.7960
2	External (pred_r^2)	0.9967	0.9973	0.9978

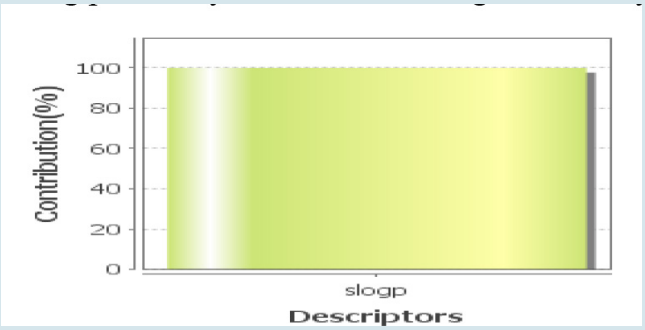


Figure 3: Contribution plot for model 1st by MLR method.

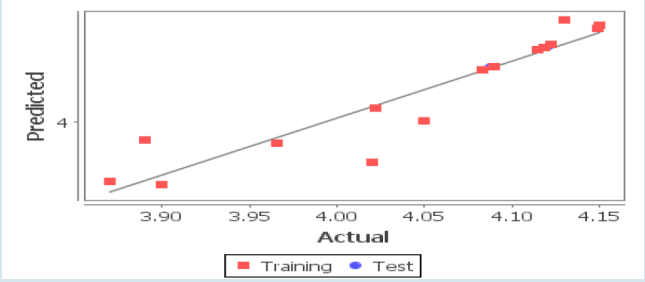


Figure 4: Fitness plot for 2D QSAR models.

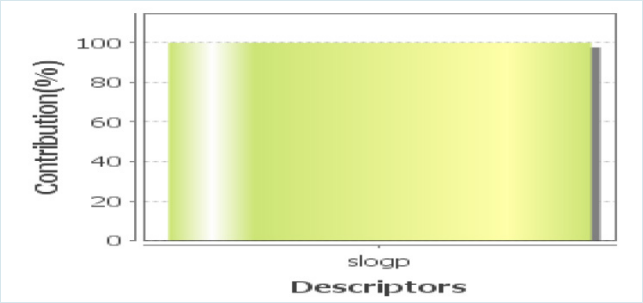


Figure 5: Contribution plot for model 2nd(PLS).

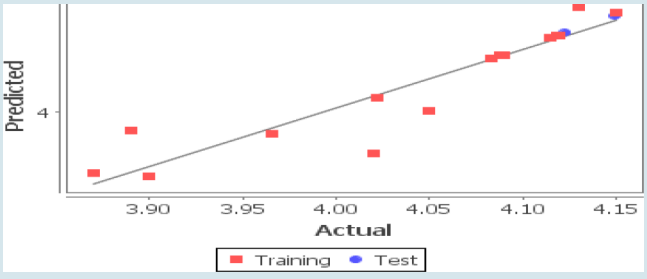


Figure 6: Fitness plot for model 2nd (PLS).

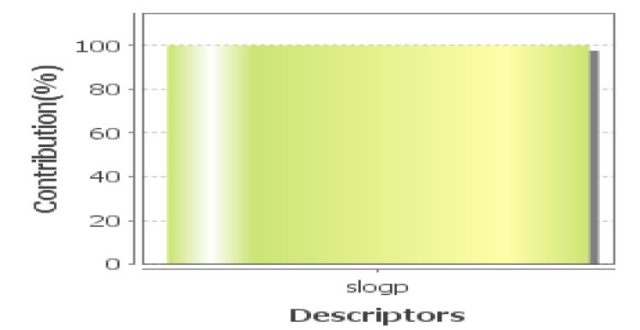


Figure 7: Contribution plot for model 3rd (PCR).

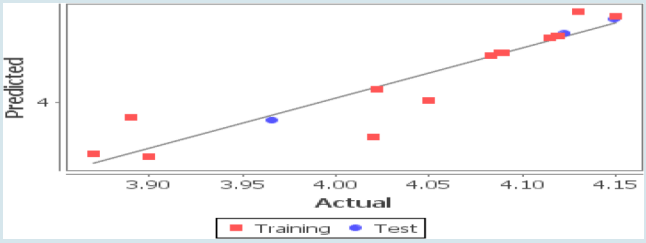
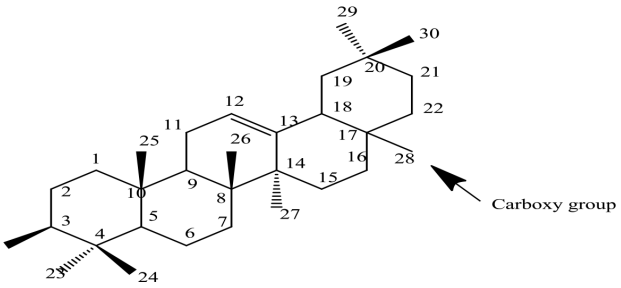


Figure 8: Fitness plot for model 3rd (PCR).

tor responsible for nematicidal activity for three models. The overall degree of prediction of descriptor was found to be around 100% in case of all three models (MLR, PLS PCR). Among the three (MLR, PLS, PCR) 2D-QSAR models, result of Multiple Linear Regression (MLR) analysis showed significant predictive power and reliability as compared to other two methods.



The results obtained from this 2D-QSAR study are in agreement with the observed SAR of *Pulsatilla koreana* saponin studied.

Hence, the model proposed in this work is useful and can be employed to design new saponin derivative of *Pulsatilla koreana* as most active nematicidalagent.

ACKNOWLEDGEMENT

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

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

QSAR: Quantitative Structure-Activity Relationship; **MLR:** Multiple Linear Regressions; **PKE:** *Pulsatilla koreana*; **N:** Optimum number of components; r^2**:** Square of Correlation Coefficient; q2: Cross-Validated Correlation Coefficient; **pred_**r^2**:** r^2 for external test set; Fischer's value: F-test.

REFERENCES

1. Hostettmann KA, *et al.* Saponins, Chemistry and Pharmacology of natural products, Cambridge university press, Cambridge, UK. 1995;4540-4.
2. Sparge SG, *et al.* Biological activity and distribution of plant saponins. J Ethnopharmacol. 2004;94(2-3):219-43.
3. Mert-Turk F. Saponin versus plant fungal pathogens. CanakkaleOnsekiz Mart University, Turkey. 2005;13-7.
4. Eva M. Extraction, Isolation and Structure Elucidation of Saponins from *Herniaria incana*. 2013;11.
5. Young HK, *et al.* Isolation of Nematicidal Triterpenoid Saponins from *Pulsatillakoreana* Root and their Activities against *Meloidogyne* incognita. Molecules. 2013;18(5):5306-16.
6. Akhtar M. Current options in integrated management of plantparasitic nematodes. Integ Pest Manag Rev. 1997;2(4):187-97.
7. Chitwood DJ. Phytochemical based strategies for nematode control. Annu Rev Phytopathol. 2002;40(1):221-49.
8. Faizi S. Isolation of nematicidal compounds from *Tagetes patula* L. yellow flowers: Structure-activity relationship studies against cyst nematode Heteroderazeae infective stage larvae. J Agric Food Chem. 2011;59(17):9080-93.
9. Seong CB, *et al.* Antitumor Activity of *Pulsatilla koreana* Saponins and Their Structure Activity Relationship. Chem Pharm Bull. 2005;53(11):1451-4.
10. V-Life Product Documentation Tutorial: QSAR, Obtaining conventional QSAR models using VLifeMDS. 1-27.
11. Malleshappa NN, *et al.* 2D QSAR studies on a series of 4-anilino Quinazoline derivatives as tyrosine Kinase (EGFR) inhibitor: an approach to design anticancer agents. Digest Journal of Nanomaterials and Biostructures. 2010;5(2):387-401.

PICTORIAL ABSTRACT