EXCLI Journal 2016;15:38-53 – ISSN 1611-2156
Received: December 07, 2015, accepted: January 05, 2016, published: January18, 2016

Original article:

QSAR STUDY OF HCV NS5B POLYMERASE INHIBITORS USING THE GENETIC ALGORITHM-MULTIPLE LINEAR REGRESSION (GA-MLR)

Hamid Rafiei[1], Marziyeh Khanzadeh[2], Shahla Mozaffari[2], Mohammad Hassan Bostanifar[1], Zhila Mohajeri Avval[2], Reza Aalizadeh[3], Eslam Pourbasheer[2] *

[1] Department of Chemistry, Dashtestan Branch, Islamic Azad University, Dashtestan, Iran
[2] Department of Chemistry, Payame Noor University (PNU), P. O. Box 19395-3697, Tehran, Iran
[3] Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

* Corresponding author: Tel: +98-45-33519448, Fax: +98-45-33519448,
e-mail: pourbasheer@ut.ac.ir

http://dx.doi.org/10.17179/excli2015-731

ABSTRACT

Quantitative structure–activity relationship (QSAR) study has been employed for predicting the inhibitory activities of the ***Hepatitis C virus (HCV) NS5B polymerase inhibitors***. A data set consisted of 72 compounds was selected, and then different types of molecular descriptors were calculated. The whole data set was split into a training set (80 % of the dataset) and a test set (20 % of the dataset) using principle component analysis. The stepwise (SW) and the genetic algorithm (GA) techniques were used as variable selection tools. Multiple linear regression method was then used to linearly correlate the selected descriptors with inhibitory activities. Several validation technique including leave-one-out and leave-group-out cross-validation, Y-randomization method were used to evaluate the internal capability of the derived models. The external prediction ability of the derived models was further analyzed using modified r^2, concordance correlation coefficient values and Golbraikh and Tropsha acceptable model criteria's. Based on the derived results (GA-MLR), some new insights toward molecular structural requirements for obtaining better inhibitory activity were obtained.

Keywords: QSAR, Genetic algorithms, Multiple linear regression, HCV

INTRODUCTION

Hepatitis C virus (HCV), identified in 1989 as the etiological agent of parenteral non-A non-B hepatitis, often causes the development of malignant chronic disease, including liver cirrhosis and hepatocellular carcinoma, frequently resulting in death (Alter et al., 1992; Choo et al., 1989; Leyssen et al., 2000). With an estimated 3 % of the global population infected with HCV, including 4.1 million in the United States alone, and no protective vaccine available at present, this disease has emerged as a serious global health problem (Wasley and Alter, 2000; Alter et al., 1999). Although significant advances have been made in the development of treatments for chronic hepatitis C, their efficacy is not universal and only 50 % success has been reported in achieving a sustained viral response for the current combination therapy with new pegylated (PEG) forms of interferon plus ribavirin (Dillon, 2004; Hügle and Cerny, 2003; Walker et al., 2003; Wang and Heinz, 2000). Moreover, this therapy has considerable liabilities including serious adverse side effects and high cost, thus highlighting the need to develop improved therapeutic options to target HCV infections (Cornberg et al., 2003).

HCV is an envelope positive-stranded RNA virus. Its single-stranded ~9.6 kb RNA genome encodes a large polyprotein of ~3010 amino acids comprising 4 structural proteins (Core, E1, E2, and p7) and 6 non-structural proteins (NS2, -3, -4A, -4B, -5A, and -5B) (Grakoui et al., 1993; Hijikata et al., 1991; Lohmann et al., 1995). One of the NS proteins, NS5B, an RNA-dependent RNA polymerase (RdRp) is the most studied target for anti-HCV therapy as it is a crucial and unique component of the viral replication machinery (Dillon, 2004; Kaushik-Basu et al., 2007; Wang and Heinz, 2000). NS5B, a 68 kDa membrane-associated protein contains motifs shared by all RdRps in which the catalytic domain is arranged around a central cleft in an organization that resembles a right hand, with the “palm” “finger” and “thumb” subdomains common to polymerases (Bressanelli et al., 2002; Love et al., 2003). Recombinant expression of active, soluble NS5B in a variety of systems has been achieved by various C-terminal deletions between 21 and 55 amino acid residues and its biochemical properties investigated (Kaushik-Basu et al., 2007). All of these reported recombinant HCV RdRps utilize a wide range of RNAs as template *in vitro* without preference, although they do prefer certain homo-polyribonucleotides to others and their activity is stimulated by GTP under specified conditions. Many screening assays for NS5B inhibitors utilize synthetic homopolymeric templates/primers. NS5B inhibitors thus far identified by these screening procedures can be broadly classified as either nucleoside (NI) or non-nucleoside (NNI) inhibitors (Kaushik-Basu et al., 2007).

Quantitative structure-activity relationships (QSAR) studies play a key role in predicting the biological activity of new compound and provide information that is useful for molecule designing and medicinal chemistry (Karbakhsh and Sabet, 2011; Noorizadeh and Farmany, 2014). QSAR model establishes the mathematical relationship between chemical properties or activities of compounds with their various structural parameters (descriptors) such as topological, physicochemical, stereochemical or electronic indices (Pourbasheer et al., 2014b; Rathod, 2011). The most important step in building QSAR models is the selection of one or more molecular descriptors that can represent the true interpretation of molecular structure with its activity or properties (Niazi et al., 2006). Therefore, a validated QSAR model can provide valuable information, not only about the effect of fragments in molecular graph, but also it can predict the biological activities without performing any experimental efforts that the designing results are not clear. In this contribution, multiple linear regression (MLR) technique was employed to build QSAR models using the theoretical molecular descriptors selected by stepwise (SW) and genetic algorithm (GA) methods based on the training set compounds (Li et al., 2008) in order to correlate the biological activities of taken compounds with their chemical strutures.

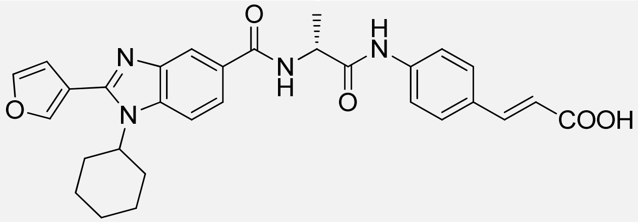
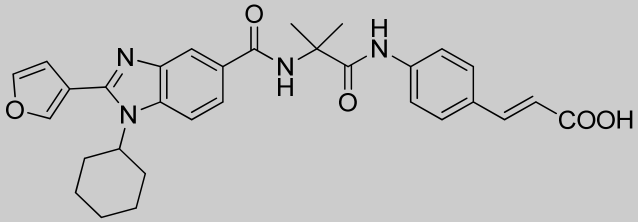
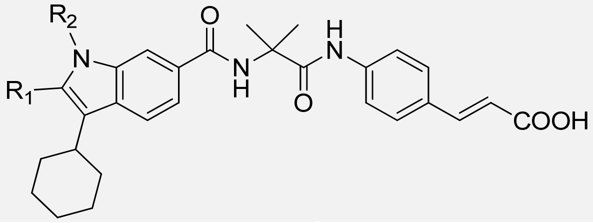
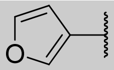
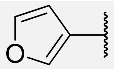
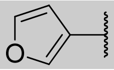
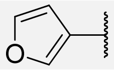
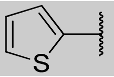
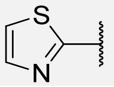
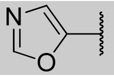
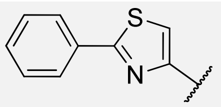
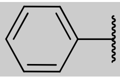
The primary goal of this work was to develop a new and validated QSAR model, and then investigating the molecular structural requirements for improving the biological activities based on the derived models.

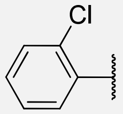
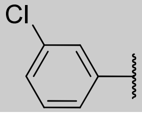
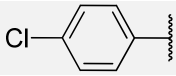
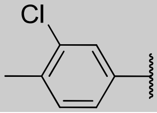
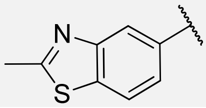
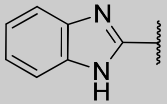
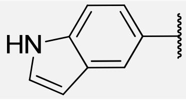
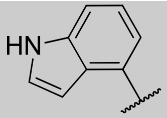
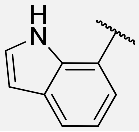
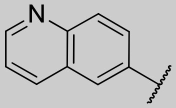
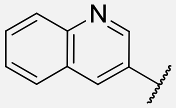
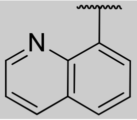
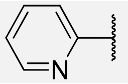
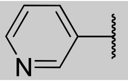
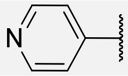
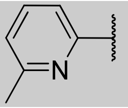
METHODOLOGY

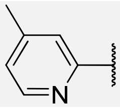
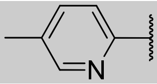
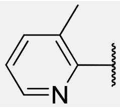
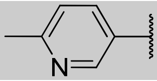
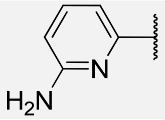
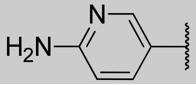
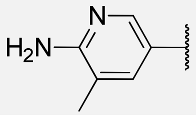
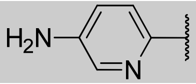
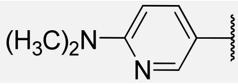
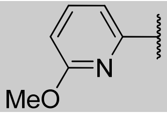
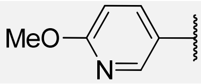
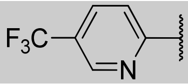
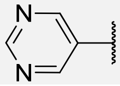
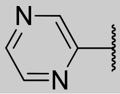
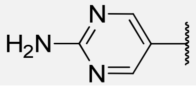
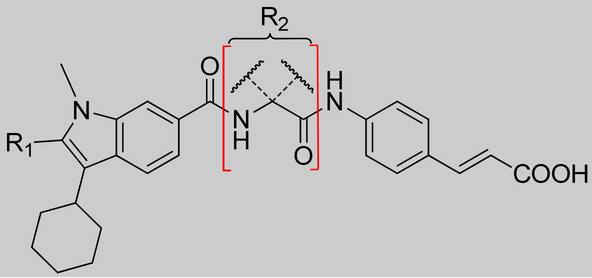
Data set

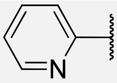
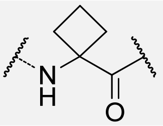
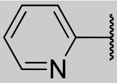
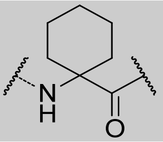
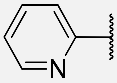
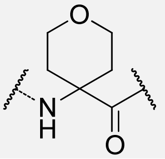
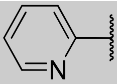
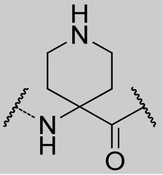
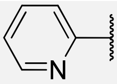
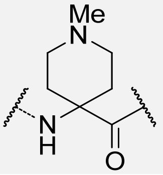
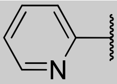
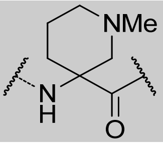
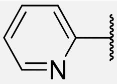
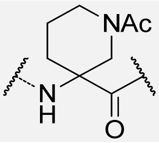
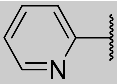
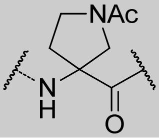
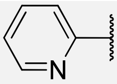
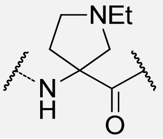
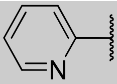
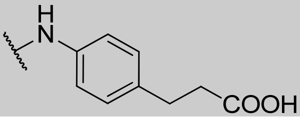
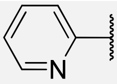
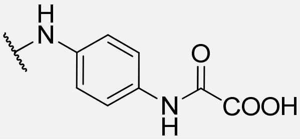
In this study, the data set consisting of 72 molecules of Indole 5-carboxamide derivatives along with their experimental inhibitory activities were taken from the literature (Beaulieu et al., 2011a, b). The chemical structures with their activities are shown in Table 1. The inhibitory activity values [IC_{50} (nM)] were converted to the logarithmic scale pIC_{50} [-log IC_{50} (M)] so as to give numerically larger value, and then used for the subsequent QSAR analyses. The molecules were divided into two subsets using principle component analysis (PCA) in which resulted in generation of the training set contained 59 compounds and the test set contained 13 compounds. The training set was employed to build the model, and the test set was used to evaluate the external prediction ability of the built models.

Table1: Chemical structures and the corresponding observed and predicted pIC$_{50}$ values by GA-MLR method

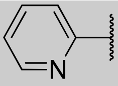
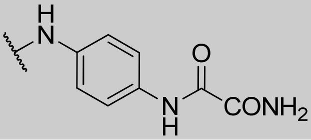
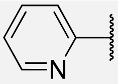
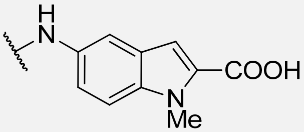
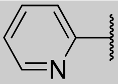
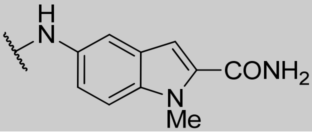
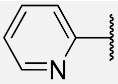
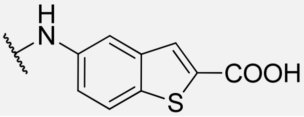
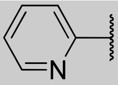
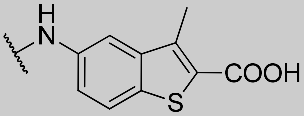
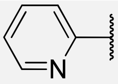
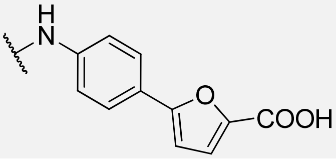
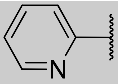
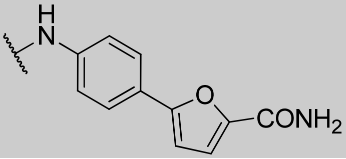
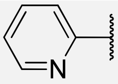
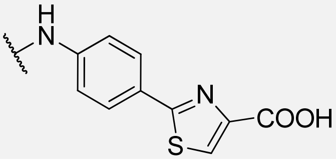
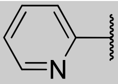
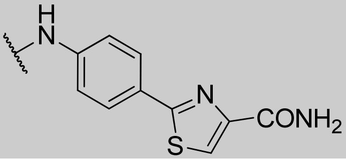
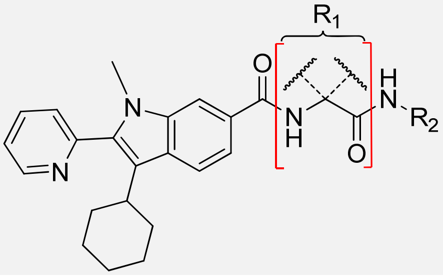
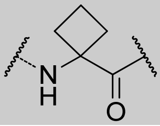
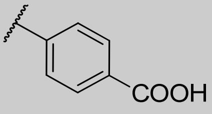
No.	**R$_1$**	**R$_2$**	**Exp. pIC$_{50}$**	**GA-MLR**
1			6.886	6.862
2			7.398	7.179
				
3		H	6.062	6.369
4		Me	6.975	6.679
5		Et	6.550	6.866
6		*i*Bu	6.206	6.198
7		Me	6.745	6.796
8		Me	7.337	7.296
9[a]		Me	7.469	6.945
10		Me	6.569	6.466
11		Me	6.462	6.757

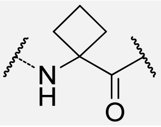
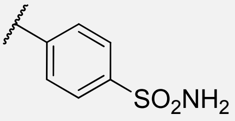
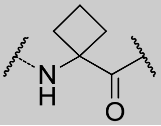
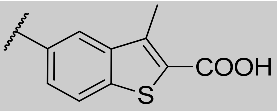
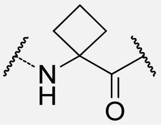
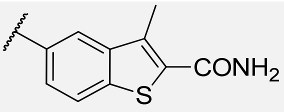
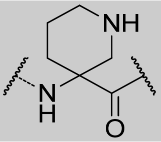
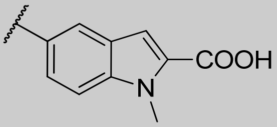
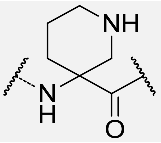
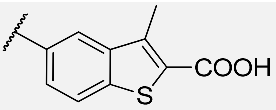
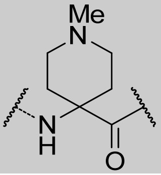
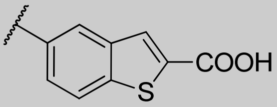
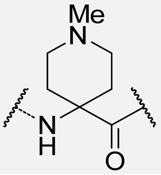
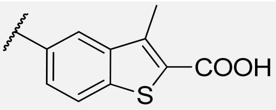
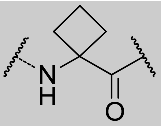
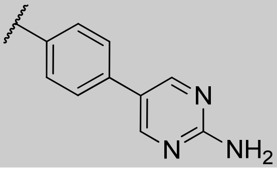
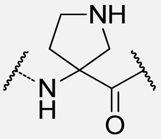
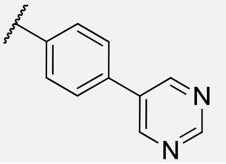
No.	R_1	R_2	Exp. pIC_{50}	GA-MLR
12		Me	6.383	6.407
13[a]		Me	6.441	6.678
14		Me	6.642	6.827
15		Me	6.526	6.569
16		Me	6.161	6.203
17		Me	6.009	6.342
18		Me	6.377	6.294
19[a]		Me	6.301	6.193
20		Me	6.357	6.411
21		Me	6.398	6.334
22		Me	6.538	6.243
23		Me	6.569	6.331
24		Me	7.444	7.327
25[a]		Me	6.963	7.077
26		Me	6.959	6.996
27		Me	7.215	6.973

No.	R_1	R_2	Exp. pIC_{50}	GA-MLR
28		Me	6.339	6.670
29		Me	7.149	6.962
30		Me	6.000	6.430
31		Me	7.081	6.865
32		Me	7.119	6.983
33		Me	6.752	6.905
34		Me	6.202	6.273
35[b]		Me	7.745	--
36[a]		Me	6.062	6.369
37		Me	7.097	6.908
38		Me	6.498	6.670
39		Me	6.804	6.568
40		Me	7.387	7.178
41		Me	7.699	7.288
42		Me	6.824	7.036
				

No.	R_1	R_2	Exp. pIC_{50}	GA-MLR
43[a]			7.284	6.974
44			6.712	7.011
45			7.167	7.623
46			7.886	7.678
47			8.000	7.724
48			7.367	7.426
49[b]			6.638	--
50			7.620	7.674
51			7.638	7.458
52			6.879	7.011
53			7.469	

EXCLI Journal 2016;15:38-53 – ISSN 1611-2156
Received: December 07, 2015, accepted: January 05, 2016, published: January18, 2016

No.	R_1	R_2	Exp. pIC_{50}	GA-MLR
54			6.907	7.445
55[a]			7.125	7.195
56			6.783	7.110
57			7.638	7.793
58			7.456	7.427
59			6.818	6.597
60			6.481	6.542
61			6.812	6.854
62[a]			6.499	6.438
				
63			7.032	6.990

No.	R_1	R_2	Exp. pIC_{50}	GA-MLR
64			6.506	6.934
65			6.914	6.749
66			7.066	6.759
67			7.357	7.154
68			7.456	7.554
69			7.770	7.843
70[a]			7.569	7.412
71			6.128	6.335
72			7.194	6.905

[a] Test set
[b] Outliers

Descriptor calculation

The two-dimensional (2D) structures of the molecules were sketched in Hyperchem v7.3 software (HyperChem, 2002) and pre-optimization was done using molecular mechanics force field (MM+) procedure, and final geometries optimization was performed using semi-empirical (AM1) method with root mean square gradient of 0.01 kcal mol^{-1}. A total of 3224 different molecular descriptors were calculated for each molecule using Dragon v5.5 package (Todeschini et al., 2010). The constant or near constant variables were removed, and then, the collinear

EXCLI Journal 2016;15:38-53 – ISSN 1611-2156
Received: December 07, 2015, accepted: January 05, 2016, published: January18, 2016

descriptors (i.e. r>0.9) were removed. The remained molecular descriptors were then taken for variable selection tool to derive the most respective subset of descriptors.

Principle Component Analysis (PCA)

The division of the dataset into training and test set is the most crucial step since based on the selected compounds, the models are being built. To divide the dataset into training and the test set, principle component analysis (PCA) (Abdi and Williams, 2010) was used so as to split the dataset based on their chemical structures diversity. The compounds in test set were selected considering the distribution in chemical structure diversity and also for avoiding the fitting problem, the better distribution of biological activities for selected compounds were considered. As a result of the PCA, 6 significant principal components (PC-s) were extracted from the variables (PC_1=49.81 %, PC_2=22.09 %, PC_3=12.25 %, PC_4=7.10 %, PC_5=6.65 %, PC_6=3.10 %,). PC_1 and PC_2 were selected for the division purpose since they covered the most variability in the dataset. The selection is first made based on the distribution of data points in PC_1 and PC_2 and then, the final candidate as test set compounds were chosen by considering the well-distribution for their biological activities.

Variable selection technique

The selection of relevant descriptors for building the predictive model is also an important step in model construction. The final goal in this step is to find the most respective descriptors which can be used to predict the biological activities with minimum error. In this contribution, we used two well-known variable selection methods including stepwise (SW) and genetic algorithm (GA). Stepwise regression includes a regression model in which the selecting of predictive variables is done by an automatic procedure (Draper and Smith, 1981) considering the F-test. Stepwise method pursues the forward selection and backward elimination rule where forward selection begins with no variable presented in the model and testing the addition of each variable improving the model outcome while, backward elimination begins with all variable and assessing the removing of variables which can improve the model by being omitted (Draper and Smith, 1981). In genetic algorithms, the initial step is creating a large number of randomly selected descriptors termed chromosome where the variables are included in each chromosome called gene (Holland, 1975; Pourbasheer et al., 2014a, c). Despite the stepwise technique, genetic algorithm is not presenting the over fitting issue, since it is using correlation coefficient of leave-one-out cross-validation (Q^2_{LOO}) as a fitness function where subset of variables are being evaluated by their fitness for selection as the most respective descriptors. Subsequently, the subsets with worse fitness function are being excluded and then, the remained subsets are breeding. Finally, the mutation is carrying out. Genetic algorithm technique was first developed by Leardi et al. (1992). Genetic algorithm and stepwise methods as selection tool were written in Matlab 6.5 program (Mathworks, 2005).

RESULTS AND DISCUSSION

The total data set was separated into a training set of 59 compounds to develop the models and a test set of 13 compounds using PCA. The training and test sets are shown in Table 1. After division of dataset, stepwise method was used to provide the most relevant descriptors for modeling purpose. Multiple linear regression method then was used to linearly correlate the selected descriptors based on the stepwise techniques on the biases of training set compounds, and then evaluated using group of compounds as test set. During the derivation of model, 2 compounds belonging to the test set were detected as outliers and excluded from analyses (Table 1). The derived linear equation based on SW-MLR is as follows:

$$pIC_{50}= 22.32 (\pm 3.511) - 4.397 (\pm 0.9607) \text{ EEig05x} + 2.673 (\pm 0.7931) \text{ GGI9} - 0.01958 (\pm 0.008726) \text{ RDF065m} - 0.7414 (\pm 0.1620) \text{ Mor19m} + 49.53 (\pm 11.34) \text{ R3u+} + 0.1809 (\pm 0.07231) \text{ C-028} \quad (1)$$

N_{train}= 59, R^2_{train}= 0.772, R^2_{test}= 0.703, R^2_{adj}= 0.745, F_{train}= 29.284, F_{test}= 0.9878, $RMSE_{train}$= 0.238, $RMSE_{test}$ = 0.265, Q^2_{LOO}=0.697, Q^2_{LGO}= 0.720, Q^2_{BOOT}= 0.712, CCC_{train}=0.871, CCC $_{test}$=0.781, r^2m=0.596, r^2m $_{average}$=0.433, MAE_{train}=0.190, MAE_{test}= 0.192.

In above equation, *N* is the number of training set compounds, R^2 is the squared correlation coefficient, *RMSE* is the root mean square error, R^2_{adj} is adjusted R^2, Q^2_{LOO}, Q^2_{LGO} and Q^2_{BOOT} are the squared cross-validation coefficients for leave one out, leave group out and bootstrapping respectively, and *F* is the Fisher *F*-statistic. CCC is concordance correlation coefficient and evaluates the degree to which pairs of observations fall on the 45° line through the origin (Pourbasheer et al., 2014d). The r^2m is modified r2 value and MAE is mean absolute error. The developed model since represented lower accuracy for test set, Golbraikh and Tropsha acceptable model criteria's was employed to investigate the reliability of the derived model (Golbraikh and Tropsha, 2002). Four conditions for accepting a model are as follows:

1. $Q^2_{LOO} > 0.5$
2. $R^2_{test} > 0.6$
3. $R_0^2 - R_0'^2/R^2 < 0.1$ and $0.85 < K' < 1.15$ or $R^2 - R_0^2/R^2 < 0.1$ and $0.85 < K < 1.15$
4. $R_0^2 - R_0'^2 < 0.3$

where R is correlation coefficient between the observed and predicted values; R_0^2 is coefficients of calculation (correlation between predicted versus observed values with intercept of zero), and $R_0'^2$ is correlation between predicted versus observed responses for regressions through the origin; K is slope and K′ is slope of regression lines through the origin. The results of this analysis were listed in Table 2. As it can be seen, the last condition for acceptance of a derived model based on SW-MLR was rejected. Therefore, the genetic algorithm as a method for variable selection was applied to the same data set (i.e. training and test set selected based on PCA) for selecting the best set of molecular descriptors. The GA-MLR analysis led to a model with six descriptors. This linear model and its statistical parameters are derived as follows:

$$pIC_{50} = 36.97\ (\pm 4.056) - 7.971\ (\pm 0.9724)\ \text{EEig05r} + 0.6368\ (\pm 0.1662)\ \text{GGI4} - 0.1752\ (\pm 0.06418)\ \text{SPAN} - 0.5972\ (\pm 0.1320)\ \text{Mor19m} + 45.88\ (\pm 13.05)\ \text{R3u+} - 5.624\ (\pm 1.617)\ \text{R5p} \quad (2)$$

N_{train}= 59, R^2_{train}= 0.792, R^2_{test}= 0.713, R^2_{adj}= 0.778, F_{train}= 32.985, F_{test}=1.3885, $RMSE_{train}$= 0.227, $RMSE_{test}$ = 0.252, Q^2_{LOO}= 0.737, Q^2_{LGO}= 0.762, Q^2_{BOOT}= 0.731, CCC_{train}=0.884, CCC_{test}=0.819, r^2m=0.666, r^2m $_{average}$=0.533, MAE_{train}=0.188, MAE_{test}= 0.213.

The PCA results were shown in Figure 1. PC_1–PC_2 loadings plot using the six descriptors for the best model (GA-MLR) were shown in Figure 2. In Figure 2, for the loadings it is confirmed that the compounds with higher biological activity values, located on the left side which are presenting a large contribution of the R3u+ descriptor, situated on the same side in Figure 1. On the other hand, compounds with lower biological activity values, on the right side, have more pronounced contributions from the other descriptors (mostly from R5p and EEig05r). Also it can be observed that the distribution of scores in Figure 1 is much more in right side and upper which represent that the most of compounds in data set have higher value for descriptors that have negative values than for the descriptors with positive effects. Therefore, the selected PCs are the true representative of the molecular descriptors that can be encoded for understanding the correlation between chemical structures and biological activities.

Table 2: Golbraikh and Tropsha acceptable model criteria's for SW-MLR and GA-MLR

	Values for SW-MLR	Values for GA-MLR	SW-MLR	GA-MLR
Condition I	0.697	0.736	Passed	Passed
Condition II	0.703	0.713	Passed	Passed
Condition III	K= 0.99612 K′= 1.0024 $R^2 - {R_0}^2/R^2$ = 0.0329 ${R_0}^2 - {R_0}'^2/R^2$ = 0.542	K= 0.99997 K′= 0.99868 $R^2 - {R_0}^2/R^2$ = 0.006 ${R_0}^2 - {R_0}'^2/R^2$ = 0.270	Passed	Passed
Condition IV	${R_0}^2 - {R_0}'^2$ = 0.358	${R_0}^2 - {R_0}'^2$ = 0.188	Failed	Passed

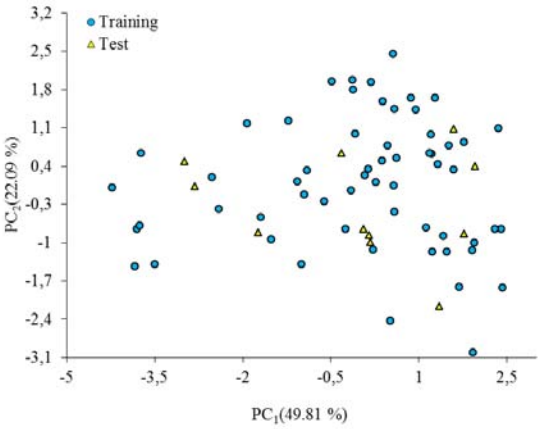


Figure 1: Principle component analysis with PC_1 and PC_2 with test set for GA-MLR result

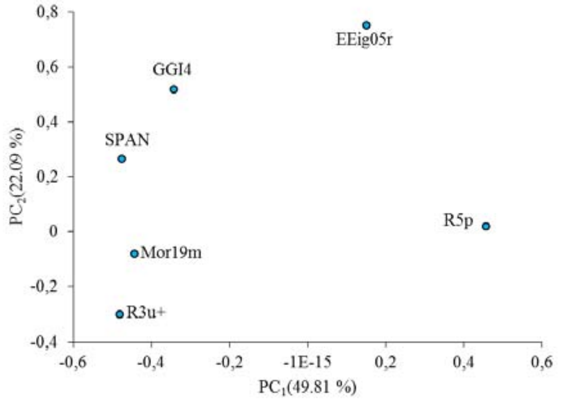


Figure 2: PC_1–PC_2 loadings plot using the six descriptors for the best model (GA-MLR)

Golbraikh and Tropsha acceptable model criteria's was employed for evaluating the prediction capability of the built GA-MLR model. The results are listed in Table 2. As it can be seen, the all conditions were accepted for GA-MLR and therefore, it was used as a main model for prediction purpose. The experimental and predicted activities based on this model were given in Table 1. The plot of the predicted pIC_{50} versus the experimental pIC_{50} is demonstrated in Figure 3. As can be seen from Table 1 and Figure 3, the calculated activity values are in good agreement with experimental activity values.

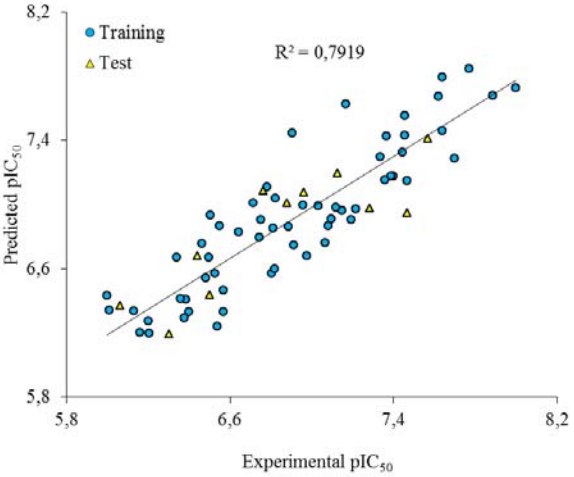


Figure 3: The predicted pIC_{50} values by the GA-MLR modeling vs. the experimental pIC_{50} values

The inter-correlation between the six selected descriptors was inspected by calculating their variance inflation factor (VIF), which are also given in Table 3. The VIF values, calculated as 1/1- r^2, where r^2 is the multiple correlation coefficient of one descriptor’s effect regressed on the remaining molecular descriptors. If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range between 1 and 5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary (Maryam et al., 2012). As it can be seen by the given information of Table 3, most of the variables had VIF values of less than 5, indicating that the GA-MLR model has statistic significance.

The built GA-MLR model was validated using the leave-one-out and leave-group-out cross-validated correlation coefficients (Q^2_{LOO} and Q^2_{LGO}). The robustness of the GA-MLR model and its predictive ability was confirmed by the high Q^2_{BOOT} source based on bootstrapping repeated 5000 times (Hadizadeh et al., 2013). The results produced by the Q^2_{LOO}, Q^2_{LGO} and Q^2_{BOOT} parameters along with other validation parameters showed the higher quality of the developed GA-MLR model. Therefore, this model can be used to predict the inhibition activity of the compounds.

The robustness of the QSAR model was further assessed by applying Y-randomization test. The dependent variable vector (inhibitory activity) was shuffled randomly and the new QSAR models (after several repetitions) would be anticipated to have low R^2 and Q^2_{LOO} values (Figure 4) (Asadollahi et al., 2011). As it can be seen from Figure 4, after 200 times shuffling the biological response for compounds, all of the derived new models were less than that of obtained in real response.

The Williams plot, the plot of the standardized residuals versus the leverage (h), is used to visualize the applicability domain (AD) of QSAR models (Vahdani and Bayat, 2011). From the Williams plot (Figure 5), it is obvious that there are only two compounds (No. 1 and No. 6 belonging to the training set) have the leverage higher than the warning $h*$ value of 0.356, thus they can be considered as structural outliers. From Figure 4, it is obvious that the standardized residuals observed for all the compounds in the training and test sets are smaller than three standard deviation units (3δ). Thus, the generated model is acceptable for prediction purpose.

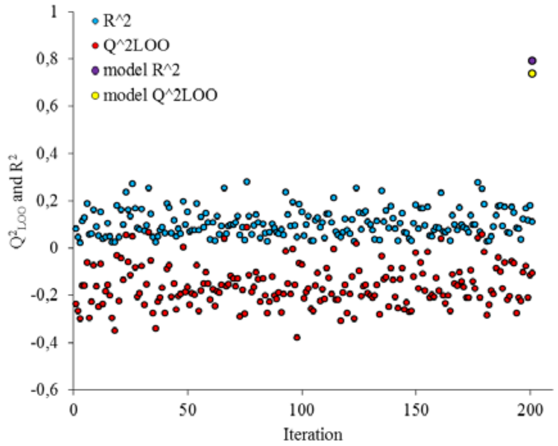


Figure 4: R^2_{train} and Q^2_{LOO} values after several Y-randomization tests for GA-MLR

Table 3: Correlation coefficient matrix of the selected descriptors with their VIF values

	EEig05r	GGI4	SPAN	Mor19m	R3u+	R5p	VIF[a]
EEig05r	1	0	0	0	0	0	1.584
GGI4	0.215	1	0	0	0	0	1.844
SPAN	-0.0051	0.513	1	0	0	0	2.449
Mor19m	-0.239	0.384	0.453	1	0	0	1.741
R3u+	-0.398	0.276	0.495	0.530	1	0	2.871
R5p	0.0655	-0.202	-0.623	-0.384	-0.685	1	2.631

[a] variance inflation factor

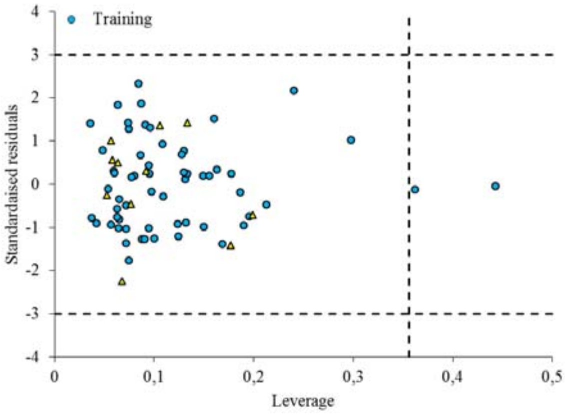


Figure 5: The William plot for the predictive GA-MLR model

Interpretation of descriptors

By interpreting the descriptors contained in GA-MLR model, some new insights can be obtained which can be helpful for understanding the correlation of chemical structure with biological activities.

The first selected descriptor is Eigenvalue 05 from edge adj. matrix weighted by resonance integrals (EEig05r) which belongs to the edge adjacency indices and encodes the connectivity between graph edges (Todeschini and Consonni, 2000). Resonance is a kind of energy stabilizing because of its delocalization effects over electrons in a bond network. As it can be seen, this descriptor represented negative effect in derived GA-MLR model encoding that increasing in the value of EEig05r by increasing the capability of the molecules (the functional groups that provide resonance in bonding with other part of bonding network) for providing more resonances would cause to decrease the pIC$_{50}$ of compounds.

GGI4 is the second selected descriptor which is representing the topological charge index of order 4 (Todeschini and Consonni, 2008). Topological charge indices are evaluating the charge transfer between atoms. These types of descriptors were first introduced by Galvez. In this concept a matrix called **M** was being obtained by multiplying the adjacency matrix **A** by the reciprocal square distance matrix (**D**$^{-2}$). However to prevent the division by zero, the diagonal entries of the distance matrix remain the same; the obtained matrix **M** called the Galvez matrix is then the unsymmetrical matrix $(A \times A)$, and A is the number of atoms in matrix. Based on the derived **M** matrix the charge term matrix (**CT**$_{\mathbf{ij}}$) which is the charge transfer between the pair of considered vertices can be obtained as follows:

$$CT_{ij} = \begin{cases} \delta_i & if\ i = j \\ m_{ij} - m_{ji} & if\ i \neq j \end{cases} \tag{3}$$

where m_{ij} is elements of matrix M, δ_i is vertex degree of i atom. CT_{ij} is also representing the net charge transfer between atom j and i. Hence, for each path length k, a topological charge index termed as G_k can be obtained as follows:

$$Gk = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \left| CT_{ij} \right| \cdot \delta(k; d_{ij}) \tag{4}$$

where $\delta(k; d_{ij}) = \begin{cases} 1 & if\ d_{ij} = k \\ 0 & if\ d_{ij} \neq k \end{cases}$ and d_{ij} is elements of distance matrix. Therefore, the G_k is the half-sum of all charge and indicate the total charge transfer between atoms placed at topological distance k. The positive sign of this descriptor in derived linear equation indicates that increasing the charge transfer between the pair of atoms would result in increase of the pIC$_{50}$ values, respectively.

The third selected descriptor (SPAN) is span R which belonged to geometrical size indices and represents the radius of the smallest sphere, centered on the mass, enclosing all atoms of a molecule (Todeschini and Consonni, 2009), and can be calculated as follows:

$$R = max_i(r_i) \tag{5}$$

where r_i is the distance of the *ith* atom from the center of the mass. Since this descriptor represents the negative sign in derived linear model, increasing the size of molecules by increasing the distance of specific moieties in molecules would result in decrease of the pIC$_{50}$ values.

Mor19m, the fourth selected descriptor of GA-MLR equation, 3D-MoRSE—signal

Received: December 07, 2015, accepted: January 05, 2016, published: January18, 2016

19/weighted by atomic masses, belongs to the 3D-MoRSE descriptors. This group of descriptors is subgroup of geometrical descriptors (Todeschini and Consonni, 2000). Value of this group of descriptors is dependent to 3D structure of molecule. 3D-MoRSE descriptors (3D-Molecule Representation of Structures based on electron diffraction) are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves (Soltzberg and Wilkins, 1977). This can be performed by infrared spectra simulation using a generalized scattering function. The Mor19m is associated with negative regression coefficient indicating that decreases in the corresponding 3D-MoRSE signal at scanning distance of 19 would result in increase of pIC_{50}value, namely.

The fifth and six descriptors (R3u+ and R5p, respectively) belong to the GETAWAY R-indices descriptors. GETAWAY descriptors are for geometry, topology and atomic-weights assembly. These descriptors are geometrical descriptors in which provide good position of substituents and fragments in molecule (Consonni et al., 2002). In addition, they can carry on good information on molecular size and shape. R3u+ (R maximal autocorrelation of lag 3/unweighted) related to the maximum steric contributions to molecules shape with the topological distance of 3 (Hall and Kier, 1995; Todeschini and Consonni, 2000). Since it presented a positive sign in derived linear equation, increasing in value of this descriptor will cause to increase of the activity (pIC_{50}). On the other hand, the other type of GETAWAY R-indices (i.e. R5p) which is R maximal autocorrelation of lag 5/weighted by polarizability would cause decrease in biological activity (pIC_{50}) due to its negative sign in obtained linear equation. Therefore, to obtain a good biological activity, the polarizibility of molecule should be decreased.

To conclude, it was observed that the capability of having more resonances in molecular graph is not appropriate and since most of the functional groups belonging to polar groups can represent the presence, therefore, the replacing of more polar groups should be avoided addressing to the negative effect of EEig05r and R5p descriptors. It was also seen that distance of substituents from mass center would cause negative effect on biological activities. However, a good biological activity can be presented if the charge transfer between bonding network and steric contributions to molecules shape increase.

CONCLUSION

A robust QSAR model was developed based on PCA-GA-MLR for a dataset consisting of 72 HCV NS5B polymerase inhibitors. The derived models were validated based on several validation techniques, and it was observed that GA-MLR is more accurate than the derived SW-MLR model. Based on the obtained results of GA-MLR, it was observed that the capability of having more resonances in molecular graph is not appropriate and since most of the functional groups belonging to polar groups can represent the presence, therefore, the replacing of more polar groups should be avoided addressing to the negative effect of EEig05r and R5p descriptors. It was also seen that distance of substituents from mass center would cause negative impact over biological activities. However, a good biological activity can be presented if the charge transfer between bonding network and steric contributions to molecules shape increase. In this work, the proposed models could identify and provide better insights about the chemical structure requirements for increasing the pIC_{50} values.

REFERENCES

Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Comput Stat. 2010; 2:433-59.

Alter MJ, Margolis HS, Krawczynski K, Judson FN, Mares A, Alexander WJ, et al. The natural history of community-acquired hepatitis C in the United States. N Engl J Med. 1992;327:1899-905.

Alter MJ, Kruszon-Moran D, Nainan OV, McQuillan GM, Gao F, Moyer LA, et al. The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. N Engl J Med. 1999;341:556-62.

Asadollahi T, Dadfarnia S, Shabani AMH, Ghasemi JB, Sarkhosh M. QSAR models for CXCR2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using in silico virtual screening. Molecules. 2011;16: 1928-55.

Beaulieu PL, Gillard J, Jolicoeur E, Duan J, Garneau M, Kukolj G, et al. From benzimidazole to indole-5-carboxamide Thumb Pocket I inhibitors of HCV NS5B polymerase. Part 1: Indole C-2 SAR and discovery of diamide derivatives with nanomolar potency in cell-based subgenomic replicons. Bioorg Med Chem Lett. 2011a;21:3658-63.

Beaulieu PL, Chabot C, Duan J, Garneau M, Gillard J, Jolicoeur E, et al. Indole 5-carboxamide Thumb Pocket I inhibitors of HCV NS5B polymerase with nanomolar potency in cell-based subgenomic replicons (part 2): Central amino acid linker and right-hand-side SAR studies. Bioorg Med Chem Lett. 2011b; 21:3664-70.

Bressanelli S, Tomei L, Rey FA, De Francesco R. Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. J Virol. 2002;76:3482-92.

Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989;244:359-62.

Consonni V, Todeschini R, Pavan M, Gramatica P. Structure/response correlations and similarity/ diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J Chem Inf Model. 2002;42: 693-705.

Cornberg M, Hüppe D, Wiegand J, Felten G, Wedemeyer H, Manns M. Treatment of chronic hepatitis C with PEG-interferon alpha-2b and ribavirin: 24 weeks of therapy are sufficient for HCV genotype 2 and 3. Z Gastroenterol. 2003;41:517-22.

Dillon J. What is the best treatment. J Viral Hepat. 2004;11:23-7.

Draper NR, Smith H. Applied regression analysis, 2nd ed. New York: John Wiley & Sons, Inc. 1981.

Golbraikh A, Tropsha A. Beware of q2!. J Mol Graph Model. 2002;20:269-76.

Grakoui A, Wychowski C, Lin C, Feinstone S, Rice C. Expression and identification of hepatitis C virus polyprotein cleavage products. J Virol. 1993;67: 1385-95.

Hadizadeh F, Vahdani S, Jafarpour M. Quantitative structure-activity relationship studies of 4-imidazolyl-1, 4-dihydropyridines as calcium channel blockers. Iran J Basic Med Sci. 2013;16:910-6.

Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci. 1995;35:1039-45.

Hijikata M, Kato N, Ootsuyama Y, Nakagawa M, Shimotohno K. Gene mapping of the putative structural region of the hepatitis C virus genome by in vitro processing analysis. Proc Natl Acad Sci U S A. 1991;88:5547-51.

Holland J. Adaptation in natural and artificial systems. Ann Arbor, MI: University of Michigan Press, 1975.

Hügle T, Cerny A. Current therapy and new molecular approaches to antiviral treatment and prevention of hepatitis C. Rev Med Virol. 2003;13: 361-71.

HyperChem. Molecular modeling system, 7.03 ed. Gainesville, FL: Hypercube, Inc., 2002.

Karbakhsh R, Sabet R. Application of different chemometric tools in QSAR study of azolo-adamantanes against influenza A virus. Res Pharm Sci. 2011;6:23-33.

Kaushik-Basu N, Bopda-Waffo A, Talele TT, Basu A, Chen Y, Kucukguzel SG. 4-Thiazolidinones: a novel class of hepatitis C virus NS5B polymerase inhibitors. Frontiers in bioscience: a journal and virtual library. Front Biosci. 2007;13:3857-68.

Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemometr. 1992;6: 267-81.

Leyssen P, De Clercq E, Neyts J. Perspectives for the treatment of infections with flaviviridae. Clin Microbiol Rev. 2000;13:67-82.

Li J, Lei B, Liu H, Li S, Yao X, Liu M, et al. QSAR study of malonyl-CoA decarboxylase inhibitors using GA-MLR and a new strategy of consensus modeling. J Comput Chem. 2008;29:2636-47.

Lohmann V, Koch J, Bartenschlager R. Processing pathways of the hepatitis C virus proteins. J Hepatol. 1995;24:11-9.

Love RA, Parge HE, Yu X, Hickey MJ, Diehl W, Gao J, et al. Crystallographic identification of a non-competitive inhibitor binding site on the hepatitis C virus NS5B RNA polymerase enzyme. J Virol. 2003;77:7575-81.

Maryam A, Mahmoud S, Mehdi N. QSAR study on the histamine (H3) receptor antagonists using the genetic algorithm: Multi parameter linear regression. J Serb Chem Soc. 2012; 77:639-50.

Mathworks. Genetic algorithm and direct search toolbox. User's guide. Natick, MA: The Mathworks Inc., 2005.

Niazi A, Bozorghi SJ, Shargh DN. Prediction of acidity constants of thiazolidine-4-carboxylic acid derivatives using ab initio and genetic algorithm-partial least squares. Turk J Chem. 2006;30:619-28.

Noorizadeh H, Farmany A. Theoretical prediction for the half wave reduction potential of organic molecules. Russ J Electrochem. 2014;50:579-86.

Pourbasheer E, Aalizadeh R, Ganjali M, Norouzi P, Banaei A. QSAR study of mGlu5 inhibitors by genetic algorithm-multiple linear regressions. Med Chem Res. 2014a;23:3082-91.

Pourbasheer E, Aalizadeh R, Ganjali MR, Norouzi P. QSAR study of IKKβ inhibitors by the genetic algorithm: Multiple linear regressions. Med Chem Res. 2014b;23:57-66.

Pourbasheer E, Aalizadeh R, Ganjali MR, Norouzi P, Shadmanesh J., Methenitis, C. QSAR study of Nav1.7 antagonists by multiple linear regression method based on genetic algorithm (GA-MLR). Med Chem Res. 2014c;23:2264-76.

Pourbasheer E, Aalizadeh R, Shokouhi Tabar S, Ganjali MR, Norouzi P, Shadmanesh J. 2D and 3D quantitative structure–activity relationship study of hepatitis C virus NS5B Polymerase inhibitors by comparative molecular field analysis and comparative molecular similarity indices analysis methods. J Chem Inf Model. 2014d;54:2902-14.

Rathod AK. Antifungal and antibacterial activities of imidazolylpyrimidines derivatives and their QSAR studies under conventional and microwave-assisted. Int J PharmTech Res. 2011;3:1942-51.

Soltzberg LJ, Wilkins CL. Molecular transforms: a potential tool for structure-activity studies. J Am Chem Soc. 1977;99:439-43.

Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim: Wiley-VCH, 2000.

Todeschini R, Consonni V. Handbook of molecular descriptors (p 445). Weinheim: Wiley-VCH, 2008.

Todeschini R Consonni V. Molecular descriptors for chemoinformatics (2 volumes). Weinheim: Wiley-VCH, 2009.

Todeschini R, Consonni V, Mauri A, Pavan M. DRAGON, software for the calculation of molecular descriptors, version 5.3. Milan, Italy: Talete srl, 2010.

Vahdani S, Bayat Z. A Quantitative Structure-Activity Relationship (QSAR) Study of anti-cancer drugs. Der Chemica Sinica. 2011;2:235-42.

Walker MP, Appleby TC, Zhong W, Lau J, Hong Z. Hepatitis C virus therapies: current treatments, targets and future perspectives. Antivir Chem Chemother. 2003;14:1-22.

Wang QM, Heinz BA. Recent advances in prevention and treatment of hepatitis C virus infections. Prog Drug Res 2000;55:1-32.

Wasley A, Alter MJ. Epidemiology of hepatitis C: geographic differences and temporal trends. Semin Liver Dis. 2000;20:1-16.