Albanian Journal of Pharmaceutical Sciences | Vol. 3 No. 1 | 2016
ISSN 2313-1772

RESEARCH ARTICLE

QSAR MODELLING OF SOME ANTICANCER PGI$_{50}$ ACTIVITY ON HL-60 CELL LINES

David Ebuka Arthur*, Adamu Uzairu, Paul Mamza, Stephen Eyije Abechi, Gideon Shallangwa

Department of Chemistry, Ahmadu Bello University (ABU) Zaria, Kaduna State, Nigeria
(*Corresponding author: hanslibs@myway.com)

Received: August 2016; Accepted: November, 2016; Published on line: November 2016.

Abstract. **QSAR (2D and 3D) studies were performed on a series of Camptothecin derivatives using Material Studio software (accelrys). QSAR study performed on 102 analogues of which 90 were used in the training set and the rest 22 considered for the test set. QSAR study performed using Genetic function approximation (GFA). GFA method came out with good correlation coefficient R^2_{train} =0.837 , cross-validated coefficient Q^2_{CV}=0.792 and R^2_{Test} of 0.9408. A highly predictive and statistically significant model was generated. The QSAR models were found to accurately predict the anticancer activity of structurally diverse test set compounds and to yield reliable clues for further optimization of the of Camptothecin derivatives in the data set.**

Keywords: Anticancer agents, Genetic Function Approximation, QSAR.

Introduction. Cancer is a major problem worldwide and is the primary cause of death in developed countries. Almost one in two men and more than one in three women in the United States will be diagnosed with cancer at some point in his or her days [1]. One of the most difficult problems arising during cancer therapy is the occurrence of cancer cell invasion responsible for the spread of tumor cells throughout the body [2]. Despite several efforts in the treatment of cancer, because of several limitations that using medications has, this disease became a big problem for the health of societies. The purpose is to develop medications with more anticancer activity and less toxicity than the present medications [3]. Computational chemistry is currently an important contributor to rational drug design [4]. The molecular and chemical computing models are used in designing new medications which resulted in saving time and cost and designing medications with more potential. Among various computational methods, QSAR has a remarkable role in designing a medication. Quantitative structure-activity relationship (QSARs) is an attempt to correlate structural or property descriptors of compounds quantitatively with biological activities. The physicochemical descriptors include parameters account for constitutional, fragment constant, thermodynamic, conformational, hydrophobicity, topology, electronic properties, steric effects, hydrogen bond-donor, hydrogen bond acceptor are determined empirically or, more recently, by computational methods [5]. QSAR models are mathematical equations which relate the chemical structure of compounds to their biological activity [6]. The fundamental idea of QSAR consists of the possibility of a relationship between a set of descriptors, which are derived from molecular structure, and a molecular response. Within this scope, several molecular descriptors, which discretely parameterize a given molecular set, have been devised. Quantum chemical calculations are thus an attractive source of new molecular descriptors, which can, in principle, express all of the electronic and geometric properties of molecules and their interactions [7]. The purpose of the present paper was to find more representative 3D descriptors able to discriminate HL-60 cell lines. The intention was also to develop QSAR models for HL-60 cell lines inhibitory effect of the aforementioned set of compounds using selected molecular descriptors and physicochemical parameters. The obtained results should give a contribution for better understanding of the binding ability of Camptothecin derivatives.

Matherial and Methods. **Data sources.** In this study, a data set of eighty-five (85) compounds from NCI database were optimized at the density functional theory (DFT) level using Becke's three-parameter Lee-Yang-Parr hybrid functional (B3LYP) in combination with the 6-31G* basis set [8, 9]. The optimized structures were employed in the generation of quantum chemical and molecular descriptors. These were then divided into training and test sets by Kennard Stone algorithm [10]. The QSAR models were generated using the Genetic Function Approximation (GFA). The GFA technique is a conglomeration of Genetic Algorithm, Friedman's multivariate adaptive regression splines (MARS) algorithm and Holland's genetic algorithm to evolve population of equations that best fit the training set data [11, 12]. **Geometry optimization** Chemical structures of the compounds were drawn using the ChemDraw software (CambridgeSoft 2010), while the molecular geometries were optimized using Spartan 14 software (Spartan 14v114) [13]at the density functional theory (DFT) level using Becke's three-parameter Lee-Yang-Parr hybrid functional (B3LYP) in combination with the 6-31G* basis set. The Spartan 14 software also resulted in the generation of a set of quantum chemical descriptors. **Descriptors calculation** The low energy conformers were then submitted for further generation of an additional set of molecular descriptors using the software "PaDel-Descriptor version 2.20" [14]. Different physicochemical descriptors were calculated for each molecule presented in Table 1. These descriptors included electronic, spatial, structural, thermodynamic and topological. This was combined to the set of quantum chemical descriptors obtained from the low energy conformer of the structures as generated by Spartan 14 Wavefunction software. **Data Pre-Treatment/Feature Selection.** It is observed that constant value and highly correlated descriptors may cause difficulties in forming QSAR models, hence the predictivity and generalization of the model fails under these conditions. In order to overcome this problem, the pre-processing for the generated molecular descriptors was done by removing descriptors having constant value and pairs of variables with correlation coefficient greater than 0.7 using "*Data Pre-Treatment GUI 1.2*" tool that uses V-WSP algorithm [15], [16]. **Dataset Division**. The dataset of eighty-five (85) molecular structures was split into training and test set by **Kennard Stone algorithm** technique using the software "Dataset Division GUI 1.2" [17]. This is an application tool used to perform rational selection of training and test set from the data set. **QSAR Model Development and Validation**. The QSAR model were developed from the training set compounds where the independent variables quantum chemical and molecular descriptors and the dependent response variable (pGI$_{50}$) were subjected to multivariate analysis by Genetic Function Approximation (GFA) technique using the material studio software. GFA was performed by using 800,000 crossovers, a smoothness value of 1.00 and other default settings for each combination. An initial of three and a maximum of five terms per equation were considered for model development. GFA measures the fitness of a model during the evolution process by calculating the Friedman lack-of-fit (LOF). **Model Validation** The developed models were validated internally by leave- one- out (LOO) cross- validation technique. In this technique, one compound is eliminated from the data set at random in each cycle and the model is built using the rest of the compounds. The model thus formed is used for predicting the activity of the eliminated compound. The process is repeated until all the compounds are eliminated once. The Cross-validated squared correlation coefficient, R^2_{cv} (Q^2) was calculated using the expression:

$$Q^2 = 1 - \frac{\sum(Y_{Obs} - Y_{Pred})^2}{\sum(Y_{Obs} - \bar{Y})^2}$$

Where Y_{obs} represents the observed activity of the training set compounds, Y_{pred} is the predicted activity of the training set compounds and Ȳ corresponds to the mean observed activity of the training set compounds. External validation was employed in order to determine the predictive capacity of the developed model as judged by its application for the prediction of test set activity values and calculation of predictive R2(R2pred) value as given by the expression:

$$R^2_{pred} = 1 - \frac{\sum\left(Y_{pred(Test)} - Y_{(Test)}\right)^2}{\sum\left(Y_{(Test)} - \bar{Y}_{(Training)}\right)^2}$$

Where $Y_{pred(Test)}$ and $Y_{(Test)}$ indicate predicted and observed activity values respectively, of the test compounds. $\bar{Y}_{(Training)}$ indicates mean activity value of the training set. R^2_{pred} is the predicted correlation coefficient calculated from the predicted activity of all the test set compounds. It has been observed that R^2pred may not be sufficient to indicated the external predictability of a model since its value is controlled by $\boldsymbol{\sum(Y_{((Test))}-\bar{Y}_{(Training)})^2}$. Thus R^2_{pred} depends on the training set mean and may not truly reflect the predictive capability of the developed model with regards to a new data set [18]. This may result in considerable numerical difference between the observed and predicted values in spite of maintaining a good overall intercorrelation.

Results and Disscussions. The total set of compounds was manually divided into a training set (90 compounds) for generating 2D-3D QSAR models and a test set (22 compounds) for validating the quality of the models. Selection of the training set and test set molecules was done on the basis of Kennard Stone algorithm technique (see section 2.5) and a wide range of activity such that the test-set molecules represent a range of biological activity similar to that of the training set; therefore, the test set is truly representative of the training set. This approach resulted in selection of compounds (see Table 1) as the test set and the remaining 90 compounds as the training set. Genetic approximation-multiple linear regression (GA-MLR) method is the standard method for multivariate data analysis. It estimates the values of the regression coefficients by applying least squares curve fitting method. For getting reliable results, dataset having typically 5 times as many data points (molecules) as independent variables (descriptors) is required. According to the statistical calculation, it was obtained the strong correlation between the topological, geometrical and functional group descriptors to the anticancer activity of the substituted analogue (leukemia cell line). The QSAR model have shown good correlation between their corresponding descriptors and biological activity. The higher the F value is, the more significant the data would be. Data would be significant if $F_{calc}/F_{table} > 1$. According to the F_{calc}/F_{table} value, it indicated that the models fits in all cases was not a chance occurrence and all models were statistically significant. Based on the R^2 criteria ($R^2 > 0.6$), the model was passed to validation step. To validate the selected prediction function, a cross-validation and an external test were carried out. Cross-validation is a practical and reliable method for testing the significance. The developed QSAR model were validated using the following statistical measures: Q^2 (coefficient of determination). A QSAR model is considered to be predictive, if the following conditions are satisfied: $Q^2 > 0.6$. The Q^2 values were used as deciding factors in selecting the optimal models. The predicted activities of the Leukemia cell line by the above model are shown in Table 2. The result of evaluation anticancer activity [predicted pGI_{50}] and correlation with anticancer activity [experiment pGI_{50}] for the model by using density functional theory (DFT) level using Becke's three-parameter Lee-Yang-Parr hybrid functional (B3LYP) in combination with the 6-31G* basis set of test set and training set can be seen at Figure 1. From the Table 1 it is evident that the predicted activities of all the compounds in the test set are in good agreement with their corresponding experimental activities and optimal fit is obtained generated by the QSARs utilizing different set of topological, geometrical and functional group descriptors. The statistically best significant model obtained by GA-MLR method with $R^2 = 0.837$ was considered, as the model showed good internal predictive power ($Q^2 = 0.792$) of 79% and predictively for the external test set ($R^2_{test} = 0.919$ for 100% data and $R^2_{test} = 0.941$) of about 90%. Consequently, QSAR model can be considered as the most suitable model for anticancer activity against leukemia cell line with both high statistical significant and excellent predictive ability. The best QSAR equation was as follows:

pGI_50 = - 1.624 **(n-Hydroxy)** + 3.150 **(AATSC7s)**
+ 4.855 **(MATS3e)** + 2.049 **(GATS3s)**
- 2.114 **(BCUTw-1h)** + 2.913 **(SpMin1_Bhv)**
+ 2.051 **(nHBint7)** - 2.440 **(minHBint7)**
+ 4.393 **(WPSA-3)** - 3.955 **(RDF145v)** - 0.941

$N_{train}=90$, $R^2_{train}=0.837$, $adjR^2_{train}=0.816$, $F_{train}=39.520$, $Q^2_{CV}=0.792$

$N_{test}= 22$ Outliers=5

Based on the coefficient of descriptor parameters involved in the QSAR model in which seen on the autocorrelation descriptor, therefore the active region of analogues can be predicted. The predictive ability of the selected model was also confirmed by Q^2_{f1}, Q^2_{f2}, R^2_m, and Concordance Correlation Coefficient (CCC) using LOO predicted values since we have separate external data set [19]. The proposed QSAR model was predictive as it satisfies the following conditions for LOO validation method: $Q^2_{f1} = 0.9213 > 0.6$, $Q^2_{f2} = 0.9158 > 0.6$, $R^2_m = 0.8867 > 0.6$ and CCC = 0.9572. The leverage values can be calculated for every compound and plotted vs. standardized residuals, and it allows a graphical detection of both the outliers and the influential chemicals in a model. Fig. 2, shows the Williams plot, the applicability domain is established inside a squared area within ±3 bound for residuals and a leverage threshold h* (h* =3p'/n), where p' is the number of model parameters and n is the number of compounds) [20]. It demonstrates that some of the compounds of the training set (ID-15, 19, 63,84 and 90) are outside of this square area while all the test set are inside of this square area. From Fig. 2, it is obvious that of the compounds of the training set (ID-15, 19, 63, 84 and 90) are outlier compounds with standard residuals >3d for the training sets. Furthermore, the chemicals have a leverage higher than the warning h* value of 0.375. **Interpretation of descriptors** The ten-variable model adequately represents the pGI_{50} data, based on direct statistics as well as validation methods. Each of the variables is a descriptor of an aspect of molecular structure and will be discussed to indicate the specific structure information encoded. By interpreting the descriptors contained in the QSAR model, it is possible to gain some insights into factors, which are related to the anticancer activity. For this reason, an acceptable interpretation of the selected descriptors is provided below. The brief descriptions of descriptors are shown in Table 2. To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor [21]. MF values are –0.053, 0.300, 0.397, 0.096, -0.041, 0.268, 0.058, -0.128, 0.138 and -0.045 for Hydroxyl, AATSC7s, MATS3e, GATS3s, BCUTw-1h, SpMin1_Bhv, nHBint7, minHBint7, WPSA-3 and RDF145v, respectively. The two 2D-descriptors, AATSC7s which correspond to Average centered Broto-Moreau autocorrelation - lag 7 / weighted by I-state, have positive mean effect (MF) which means they have positive impact on activity; therefore in the future, their values should be kept as high as possible. The high value of mean effect for MATS3e shows the significance of this descriptor in the model. MATS3e (Moran autocorrelation - lag 3 / weighted by Sanderson electronegativities) belongs to the 2D autocorrelation descriptors. The 2D autocorrelation descriptors have been successfully employed by Duchowicz et al. (2005) [22]. In these descriptors, the molecule atoms represent a set of discrete points in space, and the atomic property and function are evaluated at those points. The physico-chemical property for **MATS3e** descriptor is Sanderson electronegativities, which relate to the electronegativities of the molecule.

RESEARCH ARTICLE

Table 1: Experimental and Predicted toxicities on different leukaemia cell lines obtained with linear models based on GA-MLR technique.

ID Nr.	NAME	NSC	HL-60 (Experimental pGI_{50})	HL-60 (Predicted pGI_{50})
1	11-FORMYL-20(RS)-CAMPTOTHECIN	606172	5.9	6.477
2	11-HYDROXYMETHYL-20(RS)-CAMPTOTHECIN	606173	-	-
3	14-CHLORO-20(S)-CAMPTOTHECIN HYDRATE	643833	5.7	6.445
4	2'-DEOXY-5-FLUOROURIDINE	27640	6.2	5.427
5	3-HP	95678	6.3	6.035
6	5,6-DIHYDRO-5-AZACYTIDINE	264880	6.1	6.389
7	5-AZA-2'-DEOXYCYTIDINE	127716	4.2	3.759
8	5-AZACYTIDINE	102816	6.3[b]	6.325
9	5-HP	107392	5.4 [b]	5.462
10	7-CHLOROCAMPTOTHECIN	249910	8.0 [b]	6.337
11	9-AMINO-20-(R,S)-CAMPTOTHECIN	629971	7.9	7.118
12	ACIVICIN	163501	5.4	5.101
13	ALLOCOLCHICINE	406042	8.0	7.807
14	ALPHA-TGDR	71851	3.7	4.855
15	AMINOPTERIN DERIVATIVE1	132483	*4.0	5.869
16	AMINOPTERIN DERIVATIVE2	184692	7.5	6.679
17	AMINOPTERIN DERIVATIVE3	134033	-	-
18	AMONAFIDE	308847	5.6	5.721
19	AN ANTIFOL	623017	*8.0	7.780
20	ANTHRAPYRAZOLE DERIVATIVE	355644	7.3	7.254
21	APHIDICOLIN GLYCINATE	303812	6.5	6.000
22	ARA-C	63878	6.8	6.125
23	ASALEY	167780	6.3	6.204
24	AZQ	182986	6.5	6.708
25	BAKER'S SOLUBLE ANTIFOL	139105	6.9 [b]	5.664
26	BCNU	409962	4.8	4.005
27	BETA-TGDR	71261	5.9	4.457
28	BISANTRENE HCL	337766	7.4	8.062
29	BREQUINAR	368390	6.1	6.337
30	BUSULFAN	750	3.9	4.016
31	CAMPTOTHECIN	94600	7.9	7.157
32	CAMPTOTHECIN ANALOG	295500	7.0	7.328
33	CAMPTOTHECIN ANALOG2	606985	8.0	7.591
34	CAMPTOTHECIN ANALOG3	295501	8.0 [b]	7.121
35	CAMPTOTHECIN BUTYLGLYCINATE ESTER HYDRO-CHLORIDE	606499	7.1	7.789
36	CAMPTOTHECIN ETHYLGLYCINATE ESTER HYDRO-CHLORIDE	606497	7.3	7.802
37	CAMPTOTHECIN GLUTAMATE HCL	610459	7.7 [b]	7.429
38	CAMPTOTHECIN HEMISUCCINATE SODIUM SALT	610456	7.5	7.030
39	CAMPTOTHECIN LYSINATE HCL	610457	7.9	8.222
40	CAMPTOTHECIN PHOSPHATE	610458	7.5	7.319
41	CAMPTOTHECIN, 9-METHOXY-	176323	8.0	7.108

GRIAS Publishing Project http://www.ajphsci.com

ID Nr.	NAME	NSC	HL-60 (Experimental pGI_{50})	HL-60 (Predicted pGI_{50})
42	CAMPTOTHECIN, ACETATE	95382	6.4	7.308
43	CAMPTOTHECIN, HYDROXY-	107124	7.7[b]	6.626
44	CAMPTOTHECIN, NA SALT	100880	7.6	7.227
45	CAMPTOTHECIN,20-O-((4-(2-HYDROXYETHYL)-1-PIPERAZINO)OAC	374028	6.6	7.693
46	CAMPTOTHECIN-20-O-(N,N-DIMETHYL)GLYCINATE HCL	618939	8.0	7.969
47	CCNU	79037	4.7	4.609
48	CHLORAMBUCIL	3088	5.1	5.390
49	CHLOROZOTOCIN	178248	3.5	4.564
50	CLOMESONE	338947	3.9	3.856
51	COLCHICINE	757	7.2[b]	7.778
52	COLCHICINE DERIVATIVE	33410	8.0	7.993
53	CYANOMORPHOLINODOXORUBICIN	357704	8.1[b]	6.827
54	CYCLOCYTIDINE	145668	6.5	5.478
55	CYCLODISONE	348948	4.9	4.510
56	DAUNORUBICIN	82151	7.2	6.972
57	DEOXYDOXORUBICIN	267469	7.5	7.602
58	DIANHYDROGALACTITOL	132313	5.3	6.271
59	DICHLORALLYL LAWSONE	126771	5.7	5.962
60	DOLASTATIN 10	376128	-	-
61	DOXORUBICIN	123127	7.2[b]	7.617
62	FLUORODOPAN	73754	4.2	4.476
63	FTORAFUR (PRO-DRUG)	148958	*3.1	3.395
64	GLYCINATE	364830	7.8	6.917
65	GUANAZOLE	1895	2.9	3.013
66	HEPSULFAM	329680	4.6	4.020
67	HYCANTHONE	142982	-	-
68	HYDROXYUREA	32065	4.8	4.947
69	INOSINE GLYCODIALDEHYDE	118994	4.1[b]	5.337
70	L-ALANOSINE	153353	5.0	5.352
71	MACBECIN II	330500	7.0	6.715
72	M-AMSA	249992	7.2	7.163
73	MAYTANSINE	153858	7.9[b]	6.570
74	MELPHALAN	8806	5.6	5.480
75	MENOGARIL	269148	6.9[b]	6.625
76	METHOTREXATE	740	-	-
77	METHOTREXATE DERIVATIVE	174121	9.4	8.586
78	METHYL CCNU	95441	4.9	4.952
79	MITOMYCIN C	26980	6.6	7.370
80	MITOXANTRONE	301739	8.1	8.142
81	MITOZOLAMIDE	353451	4.5	4.512
82	MORPHOLINODOXORUBICIN	354646	8.6	8.016
83	N-(PHOSPHONOACETYL)-L-ASPARTATE (PALA)	224131	3.6	4.086

ID Nr.	NAME	NSC	HL-60 (Experimental pGI$_{50}$)	HL-60 (Predicted pGI$_{50}$)
84	N,N-DIBENZYL DAUNOMYCIN	268242	*4.7	5.183
85	NITROGEN MUSTARD	762	6.7	5.957
86	OXANTHRAZOLE	349174	6.3	6.839
87	PCNU	95466	4.2 ^{b}	4.927
88	PIPERAZINE DRUGSMAINATOR	344007	5.9 ^{b}	5.872
89	PIPERAZINEDIONE	135758	7.0	6.567
90	PIPOBROMAN	25154	4.8	4.586
91	PORFIROMYCIN	56410	6.3 ^{b}	6.559
92	PYRAZOFURIN	143095	6.3 ^{b}	7.139
93	PYRAZOLOACRIDINE	366140	6.5	6.256
94	PYRAZOLOIMIDAZOLE	51143	3.4	3.995
95	RHIZOXIN	332598	8.0	6.732
96	RUBIDAZONE	164011	7.1	7.074
97	SPIROHYDANTOIN MUSTARD	172112	4.4	5.300
98	TAXOL	125973	8.4	8.322
99	TEROXIRONE	296934	6.2 ^{b}	6.310
100	TETRAPLATIN	363812	6.2	6.259
101	THIOCOLCHICINE	361792	7.6	7.082
102	THIOGUANINE	752	5.9 ^{b}	5.855
103	THIO-TEPA	6396	5.1	5.202
104	TRIETHYLENEMELAMINE	9706	6.2	6.603
105	TRIMETREXATE	352122	7.6	7.579
106	TRITYL CYSTEINE	83265	6.3	7.135
107	URACIL NITROGEN MUSTARD	34462	5.9	4.996
108	VINBLASTINE SULFATE	49842	9.4	9.115
109	VINCRISTINE SULFATE	67574	7.0	7.615
110	VM-26	122819	7.2	6.268
111	VP-16	141540	6.0	7.138
112	YOSHI-864	102627	3.7	4.176

Where superscript **letters (b)** represent test sets for the cancer cell line, and * identifies compounds found outside the applicability domain (outliers) of the model

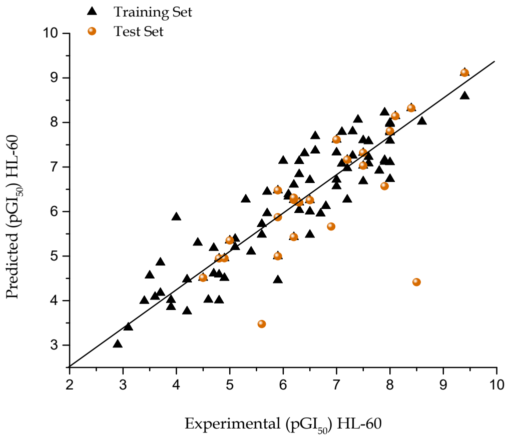

Figure 1. The predicted pGI50 against the experimental value for the training and test sets of HL-60 leukaemia cell line.

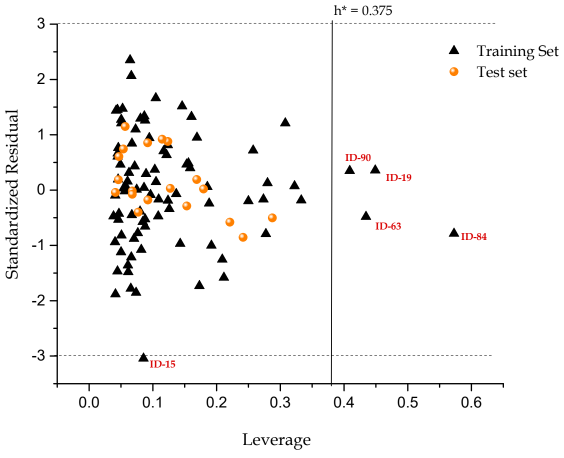

Figure 2: The Williams plot, the plot of the standardized residuals versus the activity (pGI50) leverage value for HL-60 dataset

Table 2: External Validation Result for H-60 cell line

Model biasness test	SystematicErrorResult	Absent
	R^2Test(100% data)	0.9190
	R0^2Test(100% data)	0.9180
Classical Metrics	Q2F1(100% data)	0.9213
(for 100% data)	Q2F2(100% data)	0.9158
	Scaled Avg.Rm^2(100% data)	0.8867
	Scaled DeltaRm^2(100% data)	0.0204
	CCC(100% data)	0.9572
	R^2Test(95% data)	0.9408
Classical Metric	R0^2Test(95% data)	0.9405
(after removing	Q2F1(95% data)	0.9446
5% data with	Q2F2(95% data)	0.9401

Table 3: Specification of entered descriptors in genetic algorithm multiple regression model of H-60.

Descriptors	Definition	ME
Hydroxyl	number of hydroxyl group (Fragment Counts)	-0.053
AATSC7s	Average centered Broto-Moreau autocorrelation - lag 7 / weighted by I-state	0.300
MATS3e	Moran autocorrelation - lag 3 / weighted by Sanderson electronegativities	0.397
GATS3s	Geary autocorrelation - lag 3 / weighted by I-state	0.096
BCUTw-1h	Number of low highest atom weighted BCUTS	-0.041
SpMin1_Bhv	Smallest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative van der Waals volumes	0.268
nHBint7	Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 7	0.058
minHBint7	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7	-0.128
WPSA-3	PPSA-3 * total molecular surface area / 1000	0.138
RDF145v	Radial distribution function - 145 / weighted by relative van der Waals volumes	-0.035

Therefore increasing the electronegativities of a molecule increases its MATS3e value. Mean effect of **BCUTw-1h** has the negative sign, which indicates that an increase in the weight of molecule leads to a decrease in its anticancer activity. The factor that affects the activity positively is **GATS3s**, a 2D-descriptor, which corresponds to Geary autocorrelation - lag 3 / weighted by I-state. **SpMin1_Bhv** is one of the Burden modified eigen values descriptors. The SpMin1_Bhv descriptors have been proposed as chemical structure descriptors derived from a new representation of molecular structure**.** **SpMin1_Bhv** is the Smallest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative van der Waals volumes. The **SpMin1_Bhv** mean effect has a positive sign. This sign suggests that the anticancer activity is directly related to this descriptor. Electro topological state atom type descriptor **nHBint7**, represents Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 7. This descriptor contributes positively which indicates that inhibitory activity of camptothecin derivatives will increases with Hydrogen Bonds of path length 7. Negative contribution of the **minHBint7** (Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7) indicates that inhibitory activity of camptothecin derivatives will increases with decrease of the molecular descriptors. The 3D-CPSA descriptor, **WPSA-3**, is the charged partial surface areas has positive mean effect that points out to enhance the activity, and its value should be kept as small as possible. WPSA-3 corresponds to 3D-CPSA descriptor PPSA-3 * total molecular surface area / 1000. **RDF145v** is one of the 3D-radial distribution function (RDF) descriptors which were proposed based on a radial distribution function. The radial distribution function is probability distribution to find an atom in a spherical volume of radius R. RDF descriptors are independent of the size and rotation of the entire molecule. They describe the steric hindrance or the structure/activity properties of a molecule. The RDF descriptor provides valuable information about the bond distances, ring types, planar and nonplanar systems, and atom types [23]. The descriptors used for the constructed QSAR model in this work encoded electronic, geometrical, and topological aspects of molecules. Appearances of these descriptors in the model reveal the role of electronic and steric interactions in inducing anticancer pGI_{50} activity on HL-60 cell lines.

Conclusions The model presented here, validated according to statistical criteria that are stricter than those typically used in QSAR studies, may serve as a guide for providing the structural requirements affecting the anticancer activities of camptothecin derivatives, through the identification of the most relevant selected molecular descriptors in the models. However, we applied the developed QSAR to predict some unknown structurally-related compound. We are particularly careful in validating the relationships with the Leave-One-Out Cross Validation method and by leaving some of the molecules as part of an external test set. Finally, the results presented in this work route to two different methodologies: (i) application of linear models for selecting the most relevant structural parameters, and (ii) engagement of flexible (property dependent) molecular descriptors. The good statistical parameters, stability and robustness of the model obtained, as assured by the validation tests applied over our data, indicate that these model can be used to design other camptothecin derivatives with improved anticancer activity.

References

1.Thun, M.J., et al., Lung cancer death rates in lifelong non-smokers. Journal of the National Cancer Institute, 2006. 98(10): p. 691-699.

2.Sohn, E.J., et al., EWS/FLI1 oncogene activates caspase 3 transcription and triggers apoptosis in vivo. Cancer research, 2010. 70(3): p. 1154-1163.

3.Ghanbari, Z., et al., Structure-Activity Relationship for Fe (III)-Salen-Like Complexes as Potent Anticancer Agents. The Scientific World Journal, 2014. 2014.

4.Csizmadia, I. and R. Enriz, The role of computational medicinal chemistry in the drug discovery process. 2000, ELSEVIER SCIENCE BV PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS.

5.Vaidya, A., et al., Quantitative structure-activity relationships: a novel approach of drug design and discovery. Journal of Pharmaceutical Sciences and Pharmacology, 2014. 1(3): p. 219-232.

6.Young, D., Computational chemistry: a practical guide for applying techniques to real world problems. 2004: John Wiley & Sons.

7.Salah, T., et al., In silico investigation by conceptual DFT and molecular docking of antitrypanosomal compounds for understanding cruzain inhibition. Journal of Theoretical and Computational Chemistry, 2016. 15(03): p. 1650021.

8.Benarous, N., et al., Synthesis, characterization, crystal structure and DFT study of two new polymorphs of a Schiff base (E)-2-((2,6-dichlorobenzylidene)amino)benzonitrile. Journal of Molecular Structure, 2016. 1105: p. 186-193.

9.Bauernschmitt, R. and R. Ahlrichs, Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. Chemical Physics Letters, 1996. 256(4): p. 454-464.

10.Kennard, R.W. and L.A. Stone, Computer aided design of experiments. Technometrics, 1969. 11(1): p. 137-148.
11.Deb, K., et al., A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on, 2002. 6(2): p. 182-197.

12.Leardi, R., R. Boggia, and M. Terrile, Genetic algorithms as a strategy for feature selection. J. Chemom, 1992. 6(5): p. 267-281.

13.Hehre, W.J. and W.W. Huang, Chemistry with Computation: An introduction to SPARTAN. 1995: Wavefunction, Inc.

14.Yap, C.W., PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem, 2011. 32(7): p. 1466-1474.
15.Panagos, P., et al., Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Science of the total environment, 2014. 479: p. 189-200.

16.Roy, K., S. Kar, and P. Ambure, On a simple approach for determining applicability domain of QSAR models. Chemometr Intell Lab Syst, 2015. 145: p. 22-29.

17.Barretina, J., et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012. 483(7391): p. 603-607.

18.Roy, K., S. Kar, and P. Ambure, On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems, 2015. 145: p. 22-29.

19.Chirico, N. and P. Gramatica, Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. Journal of chemical information and modeling, 2011. 51(9): p. 2320-2335.

20.Netzeva, T.I., et al., Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. ATLA, 2005. 33: p. 155-173.

21.Pourbasheer, E., et al., Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity. European journal of medicinal chemistry, 2009. 44(12): p. 5023-5028.

22.Duchowicz, P.R., et al., A new search algorithm for QSPR/QSAR theories: Normal boiling points of some organic molecules. Chemical Physics Letters, 2005. 412(4): p. 376-380.