Journal of King Saud University – Science 32 (2020) 324–331

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

Activity and toxicity modelling of some NCI selected compounds against leukemia P388ADR cell line using genetic algorithm-multiple linear regressions

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

Department of Chemistry, Ahmadu Bello University (ABU), Zaria, Kaduna State, Nigeria

ARTICLE INFO

Article history:
Received 13 February 2018
Accepted 21 May 2018
Available online 23 May 2018

Keywords:
QSAR
Multiple linear regression
Drugs
Genetic algorithm
Validation
Molecular descriptors

ABSTRACT

Cancer-causing nature is one of the toxicological endpoints bringing about the most elevated concern. Likewise, the standard bioassays in rodents used to survey the cancer-mitigating capability of chemicals and medications are expensive and require the sacrifice of animals. Thus, we have endeavored the development of a worldwide QSAR model utilizing an information set of 85 compounds, including drugs for their anti-leukemia potential. Considering expansive number of information focuses with different structural elements utilized for model development (ntraining = 68) and model validation (ntest = 17), the model developed in this study has an encouraging statistical quality (leave-one-out Q2 = 0.833, R2pred = 0.716) for pLC50 and (leave-one-out Q2 = 0.744, R2pred = 0.614) for pGI50. Our developed model suggests that the absence of methanal fragments, low dipole moment and presence of some 2D autocorrelated molecular descriptors reduces the carcinogenicity. Branching, size and shape are found to be crucial factors for drug-mitigating carcinogenicity.

1. Introduction

Drugs and other chemical agents that interact with specific enzymes are usually shown as graphs and paths when establishing a relationship with their bio-activities (Speck-Planche et al., 2012b). Each vertex in the polygonal path represents a unique property referred to as molecular descriptors of a molecule. For the past decade, drug researchers have established that the geometry of drugs plays an important role in influencing their functions when complexed with a targeted receptor (Dunnington and Schmidt, 2015). This information justifies that the molecular descriptors of chemical compounds are correlated to their chemical properties, such as the large number of topological indices that have been reported for isomer discrimination and the study of molecular complexity by Arthur (Arthur et al., 2016a), others such as the rational combinatorial library design for deriving multilinear regression models were also reported (Andrada et al., 2015).

At present, cancer is one of the leading cause of death in the human population around the world, and it is predicted to increase within that trend in the coming years (Alanazi et al., 2014). The use of chemical agents to inhibit cancer cell growth is the cheapest and most promising treatment for this disease. A major advantage of chemotherapy is its use to treat a different type of cancer, where surgery and radiation therapies are limited (Rischin et al., 2000, Kashiwagi et al., 2011). The presence large libraries of discovered compounds with high activities compiled by drug databanks and institutes such as National Cancer Institute gives options of drugs that can be studied but at the same time provides a compelling problem which involves the factor time and capital cost in experimentally screening and validating the effectiveness of the new drug.

QSAR analysis is an effective method for optimizing lead compounds and designing new drugs. It is used in predicting the activity, toxicity, and carcinogenicity of compounds based on the molecular descriptors of compounds established in appropriate mathematical models. The rapid development of computational chemistry software has improved the chances and reduced the time spent in obtaining chemical parameters of compounds for this study. The aim of this research is to obtain two new models, one to predict the activity and the other toxicity of the selected dataset

* Corresponding author.
E-mail address: davidebukaarthur@gmail.com (D.E. Arthur).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2018.05.023
1018-3647/

and hopefully able to predict new strategies with improved activities capable of mitigating cancer in drug-resistant P388ADR leukaemia cell line (Gagic et al., 2016; Chen et al., 2015, Speck-Planche et al., 2012a, Zhao et al., 2013).

2. Materials and methods

The computational hardware and software used in this work includes: Computer (HP pavilion Intel(R) Core i5-4200U with 1.63 Hz and 2.3 Hz processor and Windows 8.1 operating system), Spartan 14 (Hehre and Huang, 1995), ChemBio Ultra 12.0 (Li et al., 2004, Evans, 2014), Padel-descriptor (Yap, 2011), MS Excel (Denton, 2001).

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 6-311G* basis set (Benarous et al., 2016, Bauernschmitt and Ahlrichs, 1996). The optimized structures were used to generate molecular descriptors using the paDEL program. We calculated 1875 descriptors (1444 1D, 2D descriptors, and 431 3D descriptors) molecular descriptors using the paDEL program (PaDEL-Descriptor, 2014) for example, atom-type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom- Centred Fragments, Empirical and Properties. WHIM, Petitjean shape index, count of chemical substructures identified by Laggner, while binary fingerprints and count of chemical substructures identified by Klekota, Roth and Frederick (Klekota and Roth, 2008), Dragon descriptor software (Talete, 2007, Mauri et al., 2006) was also used to calculate some other descriptors such as 3D-MoRSE descriptors, GETAWAY descriptors, WHIM descriptors and Drug-like indices. We likewise incorporate into the analyses 5 other molecular descriptors calculated from the DFT computation (dipole moment, the energy of the HOMO and LUMO molecular orbitals, total energy and HOMO–LUMO gap).

2.1. Scaling of activities and descriptors data

The response variable (biological activities) and the explanatory variable (molecular descriptors) and were scaled using auto-scaling and range scaling procedure. According to Golbraikh et al. (2003), the modeling set and the evaluation set were scaled separately. Usually, variables with larger pre-scaled value have high coefficient and those with smaller pre-scaled values has a low coefficient in the regression equation (Foudah et al., 2014). Hence, the need to transform the variables data to the standard data by subtracting the mean and dividing by its standard deviation.

$$x_i' \frac{(x_i - \hat{\mu}_i)}{\hat{\sigma}_i}$$

where x_i is the original descriptors,$\hat{\mu}_i$ is the arithmetic mean of each descriptors, $\hat{\sigma}_i$ is the standard deviation, and x_i' the scaled descriptors. This process removes the dependence of the regression coefficient on unit. This is good for cases where variables indicate concentrations or amounts of chemical compounds, or were variables represent measurements in unrelated units (Wehrens, 2011, Mevik et al., 2011).

Range scaling techniques or normalization usually give linear transformation that set the maximum and minimum of each scale to be [0, 1] or [−1, 1], etc. Here, the minimum value in a vector (a column representing a given variable "y") is subtracted from every data point "y_n" of N samples and the results are divided by the range.

$$N_{[-1,1]}(y_n) = 2\left(\frac{y_n - y_{min}}{y_{max} - y_{min}}\right) - 1$$

where y_{min} and y_{max} are, respectively, the minimum and maximum values that can be found in the data set, with respect to all the data points and the variable to normalize. The minimum and maximum value of the evaluation set was used in this normalization procedure (Tropsha, 2010, Roy et al., 2013). These range-scale descriptors have a minimums and maximum the value of −1 and 1 respectively.

These compounds were then divided into training and test sets by the Kennard-Stone algorithm (Kennard and Stone, 1969). The QSAR models were generated using the Genetic Function Approximation (GFA). The GFA technique is a collection of Genetic Algorithm used to evolve a population of equations that best fit the training set (Deb et al., 2002, Leardi et al., 1992). A unique feature of GFA is that it yields a population of models, instead of generating a single model. The developed models were then subjected to internal and external validation and Y-randomization tests so as to justify their predictability (Tropsha, 2010).

2.2. Splitting of data-set into modelling sets and evaluation test sets

The data set was divided into two sets, the modelling set, and test set. The modelling set is used in developing the model, it contains eighty percent of the entire data set. While the test set which constitutes the remaining twenty percent of the whole data set was not used in the construction of the model but to ascertain the predictive ability of the model (Tropsha, 2010).

2.2.1. Model development

Multiple Linear Regression was used to establish a relationship between the bioactivities (pGI_{50}) and the molecular descriptors. The model was written such that sum-of-squares difference between the experimental and predicted values of the bioactivities were minimized.

2.2.2. Evaluation of the QSAR model

The QSAR models developed was validated by subjecting the models to some statistical tests as: $R^{2:}$ Fishers test; cross-validated test and pred R^2.

2.2.3. Validation of the QSAR model

The ability of a QSAR equation to predict the bioactivity of the compounds within the training set was carried out, using the leave-one-out cross-validation method. The cross-validation regression coefficient ($Q_{cv}{}^2$) is given as:

$$q^2_{CV} = 1 - PRESS/TOTAL = 1 - \frac{\sum_{i=1}^{n}(y_{exp} - y_{pred})^2}{\sum_{i=1}^{n}(y_{exp} - \bar{y})^2}$$

where y_{pred}, y_{exp}, and y~ are the predicted, experimental, and mean values of experimental activity respectively. It has been reported that high estimation of statistical attributes is not enough to justify the ability of a model, and so to assess the predictive capacity of the new QSAR model, the method depicted by Golbraikh and Tropsha (2002) and Roy et al. (2015) were utilized. The coefficient of determination for the test set R^2_{test}, was calculated through the accompanying mathematical statement

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

$\bar{Y}_T raining$ is the average activity value of the training set compounds (Tropsha et al., 2003). Additional assessment of the predictive power of the QSAR model for the test set compounds was done by calculating the value of ($\mathbf{r^2_m}$), using the rm^2 metric by Roy et al. (2013).

3. Results and discussion

The predicted activities of the training set compounds which was generated by the Material Studio Software, as well as the predicted test set values calculated using MSExcel 2013 (Carlberg, 2014) are presented in Table 1.

The results for the validation of the QSAR models presented as

Table 1
Chemical Names of Dataset with NSC numbers with Activity/Toxicity value.

Serial Number (ID)	NAME	NSC	P388ADR (Experimental pGI_{50})	P388ADR (Predicted pGI_{50})	P388ADR (Experimental pLC_{50})	P388ADR (Predicted pLC_{50})
1	2′-DEOXY-5-FLUOROURIDINE	27,640	6.6	4.800[a]	3	3.122
2	3-HP	95,678	6.4	6.017	3	3.246
3	5,6-DIHYDRO-5-AZACYTIDINE	264,880	5	5.228	2.8	2.962
4	5-AZA-2′-DEOXYCYTIDINE	127,716	7	6.325	3.5	2.975
5	5-AZACYTIDINE	102,816	6.5	5.812	2.7	2.969
6	5-HP	107,392	5.9	6.653	2.8	2.920
7	ACIVICIN	163,501	6.5	6.823	3	2.966
8	ALPHA-TGDR	71,851	4.1	6.147	2.3	2.278
9	AMINOPTERIN DERIVATIVE1	132,483	7.3	7.411	4	4.339
10	AMINOPTERIN DERIVATIVE2	184,692	7.7	9.171[a]	4	4.016
11	AMINOPTERIN DERIVATIVE3	134,033	8	7.602	4	3.867
12	AMONAFIDE	308,847	5.9	5.679	3.9	4.158
13	AN ANTIFOL	623,017	8	7.770	4	3.871
14	ANTHRAPYRAZOLE DERIVATIVE	355,644	6.7	5.841	4	3.937[b]
15	ARA-C	63,878	7.3	6.608	4	2.982[b]
16	ASALEY	167,780	5.8	7.769[a]	4.1	3.824[b]
17	AZQ	182,986	5.6	4.377	3.9	3.756
18	BAKER'S SOLUBLE ANTIFOL	139,105	5.6	5.924	3	3.035
19	BCNU	409,962	5.1	4.818	3.5	3.006[b]
20	BETA-TGDR	71,261	6.6	6.225	2.9	2.725
21	BISANTRENE HCl	337,766	4.1	5.612	3.6	3.904
22	BREQUINAR	368,390	6.7	7.403[a]	3.3	3.237
23	BUSULFAN	750	4	4.105	3.6	3.501
24	CAMPTOTHECIN	94,600	7.6	7.603[a]	4.5	3.992
25	CAMPTOTHECIN, HYDROXY-	107,124	7.4	7.499	4.2	3.810[b]
26	CAMPTOTHECIN, NA SALT	100,880	7.5	7.508	3.8	3.806
27	CCNU	79,037	5.3	5.723	3.7	3.431
28	CHLORAMBUCIL	3088	5.2	5.396	3.3	3.545
29	CHLOROZOTOCIN	178,248	4.1	4.861	2.9	3.213[b]
30	CLOMESONE	338,947	4.6	4.845	2.3	2.520
31	COLCHICINE	757	5.9	5.710	3.2	3.557
32	CYANOMORPHOLINODOXORUBICIN	357,704	8.6	5.437[a]	4.6	4.032
33	CYCLOCYTIDINE	145,668	6.9	6.649	3	3.278[b]
34	CYCLODISONE	348,948	5.1	5.175	2.7	2.553
35	DAUNORUBICIN	82,151	5.9	6.556	4	3.762[b]
36	DEOXYDOXORUBICIN	267,469	6.4	5.659[a]	3.8	3.795
37	DIANHYDROGALACTITOL	132,313	5.8	6.496[a]	3.8	3.797
38	DICHLORALLYL LAWSONE	126,771	5.8	7.239[a]	3.7	3.754
39	FLUORODOPAN	73,754	3.5	4.617[a]	2.6	2.345
40	FTORAFUR (PRO-DRUG)	148,958	4.6	5.204	3	2.593
41	GUANAZOLE	1895	3	3.224	2	1.793
42	HEPSULFAM	329,680	4.1	3.962	2.6	3.425[b]
43	HYCANTHONE	142,982	5.2	5.990	4.1	4.019
44	HYDROXYUREA	32,065	4.2	4.411	2.7	2.826
45	INOSINE GLYCODIALDEHYDE	118,994	4.2	5.417[a]	2.6	3.808[b]
46	L-ALANOSINE	153,353	5.1	5.059	3.3	3.268
47	M-AMSA	249,992	6.6	6.268	4.1	4.138
48	MAYTANSINE	153,858	8	7.321	4.6	4.535
49	MELPHALAN	8806	5.2	5.636	3.7	3.621
50	MENOGARIL	269,148	5.9	6.745	4.3	3.884
51	METHOTREXATE	740	7.6	8.522[a]	4.1	3.908[b]
52	METHYL CCNU	95,441	5.8	5.748	3.6	3.458
53	MITOMYCIN C	26,980	5.9	5.445	4.6	3.559
54	MITOXANTRONE	301,739	7.7	6.715	4.6	4.055
55	MITOZOLAMIDE	353,451	4.9	5.273	2.9	3.080
56	MORPHOLINODOXORUBICIN	354,646	8.6	7.223[a]	4.7	4.366[b]
57	N-(PHOSPHONOACETYL)-L-ASPARTATE (PALA)	224,131	4.1	3.986	2	2.090[b]
58	*N,N*-DIBENZYL DAUNOMYCIN	268,242	5.8	6.780	4.3	3.855
59	NITROGEN MUSTARD	762	7.2	6.971	4.1	3.473
60	OXANTHRAZOLE	349,174	5.9	5.930	3.6	4.128[b]
61	PCNU	95,466	4.6	5.214	2.9	3.380
62	PIPERAZINE DRUGSMAINATOR	344,007	4.6	5.154[a]	3	3.372
63	PIPERAZINEDIONE	135,758	6.6	6.145[a]	3	3.135
64	PIPOBROMAN	25,154	4.8	4.493	3.4	3.363
65	PORFIROMYCIN	56,410	5.1	5.254	3	3.299

Table 1 (*continued*)

Serial Number (ID)	NAME	NSC	P388ADR (Experimental pGI_{50})	P388ADR (Predicted pGI_{50})	P388ADR (Experimental pLC_{50})	P388ADR (Predicted pLC_{50})
66	PYRAZOFURIN	143,095	6.3	6.013	2.3	2.905
67	PYRAZOLOACRIDINE	366,140	6.7	5.772	4.6	4.167
68	PYRAZOLOIMIDAZOLE	51,143	3.5	3.786	2	2.640
69	RHIZOXIN	332,598	8	7.654	4.7	4.825
70	RUBIDAZONE	164,011	5.3	6.821	3.9	3.646
71	SPIROHYDANTOIN MUSTARD	172,112	4.5	5.463	3.6	3.338[b]
72	TAXOL	125,973	6.2	5.974	4	5.186
73	TEROXIRONE	296,934	5.7	4.698[a]	2.6	2.477
74	THIOPURINE	755	6	5.405	3.8	3.690
75	THIOGUANINE	752	6.7	5.517[a]	3.1	2.871[b]
76	THIO-TEPA	6396	5.1	4.822	3.1	2.401[b]
77	TRIETHYLENEMELAMINE	9706	6.3	5.981	1.1	1.329
78	TRIMETREXATE	352,122	7.6	7.571	3.6	3.928
79	TRITYL CYSTEINE	83,265	5.3	4.751	3.9	4.015
80	URACIL NITROGEN MUSTARD	34,462	5.8	5.530	3.5	3.461
81	VINBLASTINE SULFATE	49,842	7	7.386	5.7	5.554
82	VINCRISTINE SULFATE	67,574	6.8	6.593	3.3	3.810
83	VM-26	122,819	6.2	4.681	4.6	4.047
84	VP-16	141,540	4.2	5.275	3.1	3.577
85	YOSHI-864	102,627	3.4	4.978	2	2.134

Where superscript a and b represent test sets for P388ADR leukemia cell line for the activity and toxicity model respectively.

Toxicity

$$pLC_{50} = -2.05459(\boldsymbol{Methanal}) - 1.70276(\boldsymbol{Shadowlength} : \boldsymbol{LX}) - 2.69649(\boldsymbol{Dipole}) - 2.48671(\boldsymbol{AATSC4i}) + 1.776385(\boldsymbol{MATS3e}) + 2.834532(\boldsymbol{SpMax\_Dt}) + 1.02878(\boldsymbol{naaaC}) + 1.55842(\boldsymbol{nAtomLAC}) + 1.070197(\boldsymbol{JGI8}) + 3.002073 \quad (1)$$

$$N_{train} = 68, R^2_{train} = 0.813, adjR^2_{train} = 0.784, F_{train} = 28.104, Q^2_{CV} = 0.8135, Q^2_{LOO} = 0.8334,$$

$$N_{test} = 17, R^2_{test} = 0.716, RMSE_{train} = 0.348, RMSE_{test} = 0.588$$

Activity

$$pGI_{50} = -2.764(\boldsymbol{AATS8e}) + 0.869(\boldsymbol{ATSC6c}) + 5.041(\boldsymbol{ATSC6i}) - 4.467(\boldsymbol{AATSC6v}) + 2.879(\boldsymbol{AATSC1p}) - 0.896(\boldsymbol{GATS7v}) + 4.927(\boldsymbol{SpMin2_{Bhs}}) + 1.413(\boldsymbol{mindsCH}) - 2.166(\boldsymbol{RDF70m}) - 0.22593$$

$$N_{train} = 68, R^2_{train} = 0.700, adjR^2_{train} = 0.654, F_{train} = 15.054, Q^2_{CV} = 0.700, Q^2_{LOO} = 0.744,$$

$$N_{test} = 17, R^2_{test} = 0.614, RMSE_{train} = 0.680, RMSE_{test} = 0.1.338$$

The calculated Q^2_{LOO} value, 0.833 and 0.744 respectively for pLC_{50} and pGI_{50} suggests a good internal validation. An external validation method where the test set constituting 30% of the dataset was subjected to the model, confirms the model was indeed good since their values were 0.716 and 0.614 respectively for the toxicity and activity models. These values suggest the robustness of the constructed models. The result of predict test set data are given in Table 1. The predicted values for pLC50 for the compounds in the training and test sets using Eq. (1) were plotted against the experimental pLC_{50} values in Fig. 1, while the for pGI_{50} it was shown Fig. 3. As can be seen from Table 1 and Figs. 1 and 3, the calculated values for the pLC_{50} as well as pGI_{50} are in good agreement with those of the experimental values.

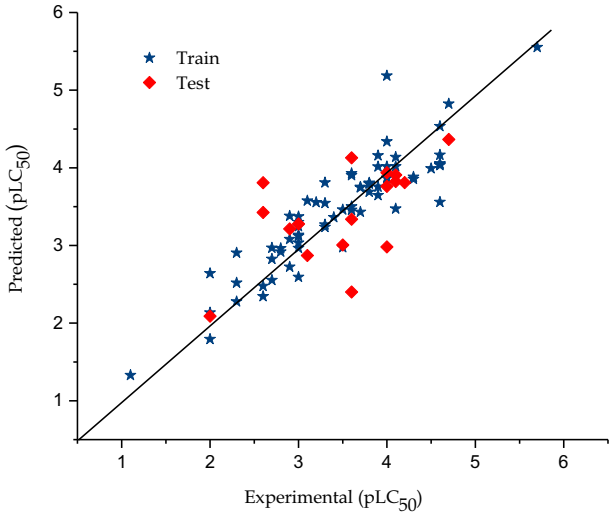

Fig. 1. Graphical representation of predicted against experimental toxicity by GA-MLR (P388ADR CELL LINE).

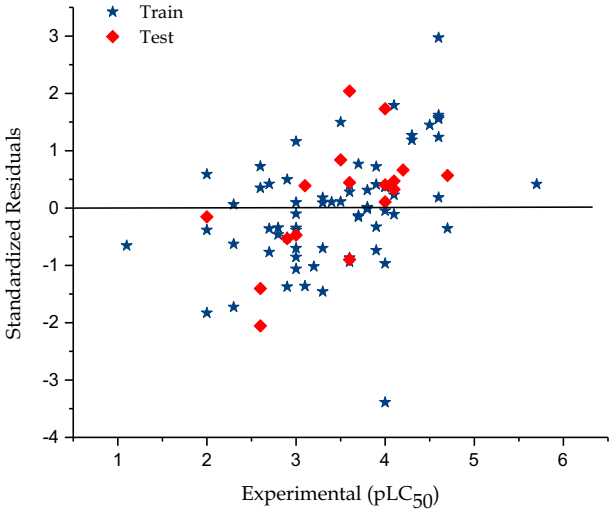

Fig. 2. Graphical representation of Standardized residual against experimental toxicity (P388ADR CELL LINE).

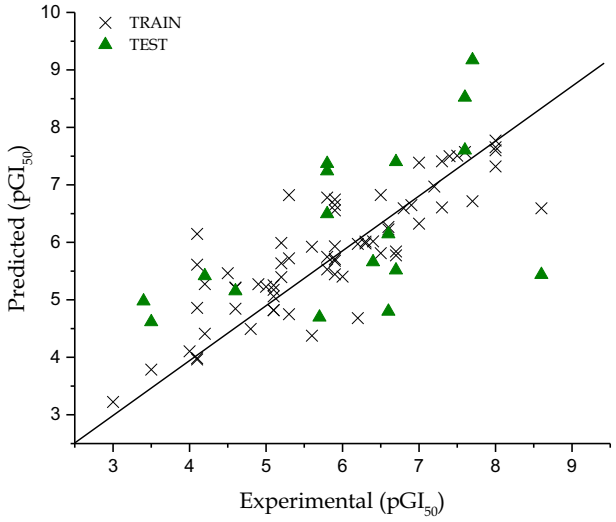

Fig. 3. Graphical representation of predicted against experimental activity by GA-MLR (P388ADR CELL LINE).

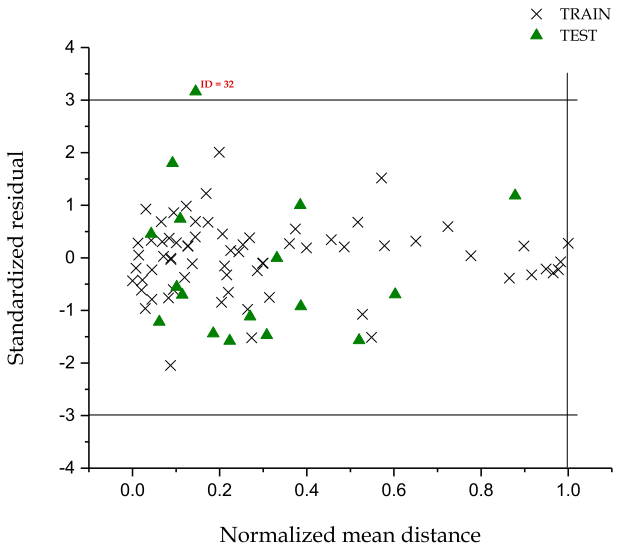

Fig. 5. Uzairu's plot: A graphical representation of Standardized residual against normalized mean distance of activity (pGI_{50}).

In 2016, similar research by Arthur et al. (2016b), published a QSAR model for the pGI_{50} and pLC_{50} model of anticancer compounds on SR leukemia cell line. The research shows that the descriptor TDB3i is the most important descriptor since it has the highest coefficient in the model, and the predicted R^2 values for the pGI_{50} (0.656) and pLC_{50} (0.580) were in good comparison with the models developed in this work. Descriptors such as number of Methanal group (nMethanal) and Secondary butyl, Sum of atom-type E-State:-F (S_Sf), were found to be principally responsible for the activity nature of the compounds on SR cell lines, thereby supporting the effect of Methanal group on the activity of the anticancer compounds in controlling cancer cells.

3.3. Applicability domain study

The applicability domain of the models were evaluated using Uzairu's plot which is novel applicability domain technique by Arthur et al. (2016a). This techniques involves plotting the standardized residuals of the activities and toxicities against the normalized mean distance between the values for the complete dataset of the molecular descriptors appearing in the model for both the toxicity and activity.

The plots shown in Figs. 5 and 6, indicates that all the compounds in both cases fell within the chemical space of the models. Fig. 5 shows the presence of one outlier with ID = 32, which is cyanomorpholinodoxorubicin, while in Fig. 6 the outlier Taxol was identified with ID = 72. These models were unable to predict the experimental values of these compounds because the molecular structure of the compound were completely different. We found out that these compounds were very large in size and they do not contain the primary molecular descriptor needed to predict their experimental values.

Also, the plot of the residual against the predicted values of pLC_{50} and pGI_{50} for both the training and test sets shown in Figs. 2 and 4 respectively. The model did not show any proportional and systematic error because the propagation of the residuals on both sides of zero is random.

The multi-collinearity amid the descriptors existing in the models was spotted by calculating their variation inflation factors (VIF), which can be calculated as follows:

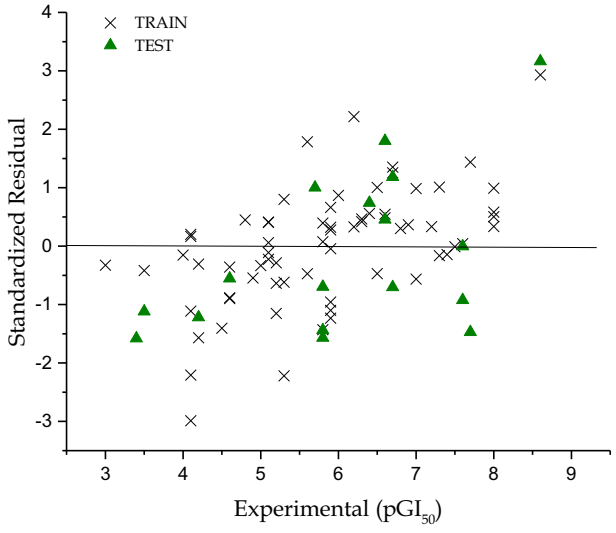

Fig. 4. Graphical representation of Standardized residual against experimental activity (P388ADR CELL LINE).

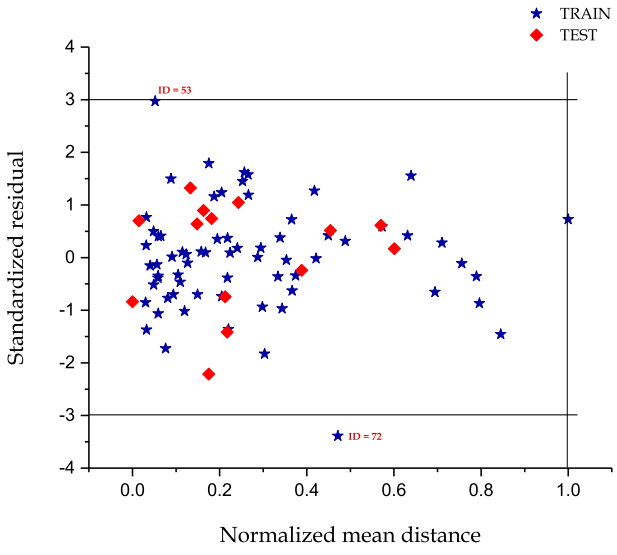

Fig. 6. Uzairu's plot: A graphical representation of Standardized residual against normalized mean distance of toxicity (pLC_{50}).

$$VIF = \frac{1}{1 - R^2}$$

where R^2 is the correlation coefficient of the multiple regression between the variables within the model (Shapiro et al., 2002). The corresponding VIF values of the descriptors were presented in Table 3.1 and 3.2. The tables show, all the variables have VIF values of less than five except for two descriptors, indicating that the descriptors were reasonably orthogonal.

In order to assess the strength of the model, the Y-randomization test was used in this study (Golbraikh et al., 2003, Tropsha et al., 2003). Y-randomization test settles whether the model is gotten through coincidental correlation, and is a true structure–activity relationship to validate the capability of the training set molecules.

The new QSAR models (after several repetitions) would be expected to have low R^2 and Q^2_{LOO} values (Table 2.1 and 2.2). If the opposite happens, then an acceptable QSAR model cannot be obtained for the specific modeling method and data. The results of Tables 2.1 and 2.2 indicate that an acceptable model is obtained by GA–MLR method and the model developed is statistically significant and robust.

3.4. Interpretation of descriptors

By interpreting the descriptors contained in the QSAR model, it is plausible to increase a few bits of information into factors, which are identified with the anti-leukemia action. Hence, a satisfactory understanding of the chosen descriptors is given below. The brief representations of descriptors shown in Table 3.1 and 3.2. To look at the comparative meaning and also the importance of every descriptor in the model, the assessment of the mean effect (MF) was established for every descriptor (Pourbasheer et al., 2009, Riahi et al., 2009); This was achieved by using an MF mathematical statement which is given as

Table 2.1
R^2_{train} and Q_{LOO} values for Toxicity model (pLC_{50}) after several Y-randomization tests.

Model	R^2_{train}	Q_{LOO}
Random 1	0.059245	0.13348
Random 2	0.125744	0.17506
Random 3	0.125553	0.30854
Random 4	0.065366	0.25758
Random 5	0.098924	0.21199
Random 6	0.159696	0.11592
Random 7	0.095924	0.1475
Random 8	0.153763	0.2116
Random 9	0.17986	0.06468
Random 10	0.061857	0.3686
Random Models Parameters		
cRp^2:	0.65333	

Table 2.2
R^2_{train} and Q_{LOO} values for Activity model (pGI_{50}) after several Y-randomization tests.

Model	R^2_{train}	Q_{LOO}
Random 1	0.068698	−0.28859
Random 2	0.068368	−0.23595
Random 3	0.130789	−0.14449
Random 4	0.067171	−0.23575
Random 5	0.214743	−0.09546
Random 6	0.138033	−0.11832
Random 7	0.180078	−0.05679
Random 8	0.104373	−0.3549
Random 9	0.174189	−0.12221
Random 10	0.051069	−0.45851
Random Models Parameters		
cRp^2:	0.515361	

Table 3.1
Definition of molecular descriptors with their corresponding Mean effects and Collinearity study for the Toxicity model.

Descriptors	Description	MF	VIF
Methanal	Functional group count	−0.09934	1.126
Shadow length: LX	Geometrical descriptor	−1.60374	3.392
Dipole (debye)	Electrostatic descriptor	−0.83386	1.138
AATSC4i	Average centered Broto-Moreau autocorrelation – lag 4/weighted by first ionization potential	−1.5074	1.129
MATS3e	Moran autocorrelation – lag 3/weighted by Sanderson electronegativities	2.163385	1.130
SpMax_Dt	Leading eigenvalue from detour matrix	1.36951	2.501
naaaC	Count of atom-type E-State:::C:	0.265287	1.292

Table 3.2
Definition of molecular descriptors with their corresponding Mean effects and Collinearity study in the Activity model.

Descriptors	Definition	MF	VIF
AATS8e	Average Broto-Moreau autocorrelation – lag 8/weighted by Sanderson electronegativities	−0.290	2.515
ATSC6c	Centered Broto-Moreau autocorrelation – lag 6/weighted by charges	0.098	1.100
ATSC6i	Centered Broto-Moreau autocorrelation – lag 6/weighted by first ionization potential	0.407	1.412
AATSC6v	Average centered Broto-Moreau autocorrelation – lag 6/weighted by van der Waals volumes	−0.159	2.184
AATSC1p	Average centered Broto-Moreau autocorrelation – lag 1/weighted by polarizabilities	0.376	1.195
GATS7v	Geary autocorrelation – lag 7/weighted by van der Waals volumes	−0.047	1.790
SpMin2_Bhs	Smallest absolute eigenvalue of Burden modified matrix – n 2/weighted by relative I-state	0.621	2.494
mindsCH	Minimum atom-type E-State: dbndCHsbnd	0.062	1.564
RDF70m	Radial distribution function - 070/weighted by relative mass	−0.056	2.117

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}}$$

MF_j is given as the mean effect for the considered molecular descriptor j, while β_j is the coefficient of the descriptor j, d_{ij} represents the values for the target descriptors of each molecule, and m is the total number of descriptors in the model. The MF values proves the relative implication of a descriptor, associated with other descriptors in the model. Its sign shows the variation direction in the estimations of the model as an effect of the descriptor values.

The dipole moment is an electric polarization descriptor; it encodes information about charge distribution in molecules. They are also important in modelling solvation properties of the compounds which depend on solute/solvent interactions since the mean effect of the dipole moment was found to be negative hence a reduction in the polarity of these compounds was found to steadily decrease the toxicity of anti-leukemia compounds. The dipole moment is given as

$$\mu = -\sum_{i=1}^{occ} \int_{(V)} \phi_i \hat{r} \phi_i d\upsilon + \sum_{a=1}^{M} Z_a \vec{R}_a$$

ϕ_i – molecular orbitals
$\hat{r}$ – electron position operator
Z_a – a-th atomic nuclear charge
$\vec{R}_a$ – position vector of a-th atomic nucleus

Methanal as a functional group count descriptor, whose mean effect was also found to negatively affect the toxicity of the compounds, while the shadow length **Lx is** the maximum dimensions of the molecular surface projections and it also negatively affects the toxicities of these compounds, this was confirmed by the negative values of the mean effect.

AATSC4i, MATS3e, and AATS8e, ATSC6C, ATSC6i, AATSC6v, AATSC1p, GATS7v are 2D Autocorrelation descriptors developed by (Todeschini and Consonni, 2009), The 2D autocorrelation descriptors have been successfully employed by Fernandez et al. (Fernandez-Lozano et al., 2015) Caballero (Caballero, 2010, Fernández et al., 2005, Vilar et al., 2009).

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 sign on the mean effects influences their behaviors in whatever model they are found in. theses descriptors as defined on Table 3.1 and 3.2 describes the weight by first ionization potential, weighted by Sanderson electronegativities, weighted by charges, by van der Waals volumes and by polarizabilities of the molecules used determines the potency of anti-leukemia compounds.

4. Conclusion

The aim of the present work was developing a QSAR study and predicting the anti-leukemia activities and toxicities of some potent NCI anticancer compounds. Different hypothetical molecular descriptors were ascertained by paDEL Software and chose by Genetic Algorithm. The developed GA–MLR model was surveyed extensively (inward and outside validations), and every one of the validation show that the QSAR model we fabricated is vigorous and agreeable. Selection of seven variables in toxicity and nine variables for the activity model showed that the descriptors methanal, shadow length LX, dipole moment, AATSC4i, MATS3e, SPMax_DT, naaaC and AATS8e, ATSC6C, ATSC6i, AATSC6v, AATSC1p, GATS7v, SpMin2_Bhs, MindsCH, RDF70m of the molecules play a main role in the anti-leukemia activity and toxicity of the compounds.

Competing interests

The authors have declared no conflict of interest.

Authors' contributions

DEA carried out the computational studies, participated in the design and drafted the manuscript. AU carried out the statistical validation of the models and participated in the write up. PM, GS and SEA participated in the design of the study and modelling the QSAR data. DEA and AU conceived of the study and coordination that helped prepare the manuscript to its final format. All authors read and approved the final manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jksus.2018.05.023.

References

Alanazi, A.M., Abdel-Aziz, A.A.M., Al-Suwaidan, I.A., Abdel-Hamide, S.G., Shawer, T. Z., El-Azab, A.S., 2014. Design, synthesis and biological evaluation of some novel substituted quinazolines as antitumor agents. Eur. J. Med. Chem. 79, 446–454.

Andrada, M.F., Vega-Hissi, E.G., Estrada, M.R., Garro Martinez, J.C., 2015. Application of k-means clustering, linear discriminant analysis and multivariate linear regression for the development of a predictive QSAR model on 5-lipoxygenase inhibitors. Chemometrics Intell. Laboratory Syst. 143, 122–129.

Arthur, D.E., Uzairu, A., Mamza, P., Abechi, S.E., Shallangwa, G., 2016a. Insilco study on the toxicity of anti-cancer compounds tested against MOLT-4 and p388 cell lines using GA-MLR technique. Beni-Suef University J. Basic Appl. Sci.

Arthur, D.E., Uzairu, A., Mamza, P., Stephen, A.E., Shallangwa, G., 2016b. Quantum modelling of the Structure-Activity and toxicity relationship studies of some potent compounds on SR leukemia cell line. Chem. Data Collect. 5, 46–61.

Bauernschmitt, R., Ahlrichs, R., 1996. Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. Chem. Phys. Lett. 256, 454–464.

Benarous, N., Cherouana, A., Aubert, E., Durand, P., Dahaoui, S., 2016. Synthesis, characterization, crystal structure and DFT study of two new polymorphs of a Schiff base (E)-2-((2,6-dichlorobenzylidene)amino)benzonitrile. J. Mol. Struct. 1105, 186–193.

Caballero, J., 2010. 3D-QSAR (CoMFA and CoMSIA) and pharmacophore (GALAHAD) studies on the differential inhibition of aldose reductase by flavonoid compounds. J. Mol. Graph. Model. 29, 363–371.

C. Carlberg 2014. Statistical Analysis: Microsoft Excel 2013, Que Publishing.

Chen, B., Zhang, T., Bond, T., Gan, Y., 2015. Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: a review of methods and resources. J. Hazard. Mater. 299, 260–279.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evolutionary Comput. 6, 182–197.

Denton, P., 2001. Generating coursework feedback for large groups of students using MS Excel and MS Word. Univ. Chem. Educ. 5, 1–8.

Dunnington, B.D., Schmidt, J.R., 2015. Molecular bonding-based descriptors for surface adsorption and reactivity. J. Catal. 324, 50–58.

Evans, D.A., 2014. History of the Harvard ChemDraw project. Angewandte Chemie Int. Ed. 53, 11140–11145.

Fernandez-Lozano, C., Cuiñas, R.F., Seoane, J.A., Fernández-Blanco, E., Dorado, J., Munteanu, C.R., 2015. Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models. J. Theor. Biol. 384, 50–58.

Fernández, M., Caballero, J., Helguera, A.M., Castro, E.A., González, M.P., 2005. Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. Bioorganic Med. Chem. 13, 3269–3277.

Foudah, A.I., Sallam, A.A., Akl, M.R., el Sayed, K.A., 2014. Optimization, pharmacophore modeling and 3D-QSAR studies of sipholanes as breast cancer migration and proliferation inhibitors. Eur. J. Med. Chem. 73, 310–324.

Gagic, Z., Nikolic, K., Ivkovic, B., Filipic, S., Agbaba, D., 2016. QSAR studies and design of new analogs of vitamin E with enhanced antiproliferative activity on MCF-7 breast cancer cells. J. Taiwan Inst. Chem. Eng.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., Tropsha, A., 2003. Rational selection of training and test sets for the development of validated QSAR models. J. Computer-Aided Mol. Des. 17, 241–253.

Golbraikh, A., Tropsha, A., 2002. Beware of q 2! J. Mol. Graph Model 20, 269–276.

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

Kashiwagi, S., Yashiro, M., Takashima, T., Aomatsu, N., Ikeda, K., Ogawa, Y., Ishikawa, T., Hirakawa, K., 2011. Advantages of adjuvant chemotherapy for patients with triple-negative breast cancer at Stage II: usefulness of prognostic markers E-cadherin and Ki67. Breast Cancer Res. 13, R122.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Klekota, J., Roth, F.P., 2008. Chemical substructures that enrich for biological activity. Bioinformatics 24, 2518–2525.

Leardi, R., Boggia, R., Terrile, M., 1992. Genetic algorithms as a strategy for feature selection. J. Chemom. 6, 267–281.

Li, Z., Wan, H., Shi, Y., Ouyang, P., 2004. Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and ChemSketch. J. Chem. Inf. Comput. Sci. 44, 1886–1890.

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon software: An easy approach to molecular descriptor calculations. Match 56, 237–248.

Mevik, B.-H., Wehrens, R., Liland, K.H., 2011. pls: Partial least squares and principal component regression. R package version 2, 3.

Pourbasheer, E., Riahi, S., Ganjali, M.R., Norouzi, P., 2009. Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity. Eur. J. Med. Chem. 44, 5023–5028.

Riahi, S., Pourbasheer, E., Ganjali, M.R., Norouzi, P., 2009. Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. J. Hazardous Mater. 166, 853–859.

Rischin, D., White, M.A., Matthews, J.P., Toner, G.C., Watty, K., Sulkowski, A.J., Clarke, J.L., Buchanan, L., 2000. A randomised crossover trial of chemotherapy in the home: patient preferences and cost analysis. Med. J. Australia 173, 125–127.

Roy, K., Chakraborty, P., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., 2013. Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. J. Comput. Chem. 34, 1071–1082.

Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemometrics Intell. Lab. Syst. 145, 22–29.

Shapiro, S., Giertsen, E., Guggenheim, B., 2002. An in vitro oral biofilm model for comparing the efficacy of antimicrobial mouthrinses. Caries Res. 36, 93–100.

Speck-Planche, A., Kleandrova, V.V., Luan, F., Cordeiro, M.N.D.S., 2012a. Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the in silico discovery of anti-breast cancer agents. Eur. J. Pharm. Sci. 47, 273–279.

Speck-Planche, A., Kleandrova, V.V., Luan, F., Cordeiro, M.N.D.S., 2012b. Rational drug design for anti-cancer chemotherapy: Multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. Bioorg. Med. Chem. 20, 4848–4855.

TALETE 2007. DRAGON. Milano, Italy.

Todeschini, R., Consonni, V. 2009. Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set), John Wiley & Sons.

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inf. 29, 476–488.

Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69–77.

Vilar, S., González-Díaz, H., Santana, L., Uriarte, E., 2009. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. J. Theor. Biol. 261, 449–458.

Wehrens, R., 2011. Chemometrics with R: multivariate data analysis in the natural sciences and life sciences. Springer Science & Business Media.

Yap, C.W., 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem 32, 1466–1474.

Zhao, L., Xiang, Y., Song, J., Zhang, Z., 2013. A novel two-step QSAR modeling work flow to predict selectivity and activity of HDAC inhibitors. Bioorg. Med. Chem. Lett. 23, 929–933.