JOTCS A
Journal of the Turkish Chemical Society
Section: A
Chemistry
TURKISH CHEMICAL SOCIETY

Toxicity Modelling of Some Active Compounds Against K562 Cancer Cell Line Using Genetic Algorithm-Multiple Linear Regressions

David Ebuka Arthur[a]*, Adamu Uzairu[a], Paul Mamza[a], Abechi Eyeji Stephen[a], Gideon Shallangwa[a]

[a]Department of Chemistry, Ahmadu Bello University Zaria, Kaduna State, Nigeria

Abstract: This research entails the modelling of the toxicity of anticancer compounds on K562 cell line, where 112 compounds that make up the data set were divided into training and test sets to be used for developing and validating the model, respectively. The internal and external validation parameter R^2 for the training and test set given respectively as 0.845 and 0.5316, justifying the robustness and the ability of the model to predict toxicity of the compounds. WPSA-3 and minHBint7 molecular descriptor is responsible for about 50% of the overall effect on the model.

Keywords: QSAR; Modeling; External Validation; Molecular descriptors; Genetic algorithm.

Submitted: October 05, 2016. **Revised:** November 17, 2016. **Accepted:** November 21, 2016.

Cite this: Arthur D, Uzairu A, Mamza P, Stephen A, Shallangwa G. Toxicity Modelling of Some Active Compounds Against K562 Cancer Cell Line Using Genetic Algorithm-Multiple Linear Regressions. JOTCSA. 2017;4(1):355–74.

DOI: To be assigned.

*Corresponding author. E-mail: hanslibs@myway.com, phone: +2348138325431.

INTRODUCTION

Drugs are often modeled as various polygonal shapes, paths, graphs, *etc.* [1]. Each vertex in the polygonal path represents an atom of the molecule, and covalent bonds between atoms are represented by edges between the corresponding vertices. As the geometry of proteins play an important role in determining the function of the protein [2], molecular descriptors of chemical compounds can be correlated to their biological activity. Presently a large number of molecular descriptors have been reported as important for the study of molecular drug design, lead optimization, and for deriving regression models [3]. At present, cancer is one of the leading disease-related cause of death of the human population in the world, and it is predicted to continue to become the leading cause of death within the coming years [4]. The use of chemical compounds to inhibit cancer cell growth, is a mainstay in the treatment of malignancies. A major advantage of chemotherapy is its ability to treat metastatic cancers, whereas surgery and radiation therapies are limited to treating cancers that are confined to specific areas. Chemotherapy has aroused many researchers' interests and a great deal of current efforts has been focusing on the design and development of different anticancer drugs. The large and extensive library of discovered compounds with high activities have been compiled by drug databanks and institutes such as National Cancer Institute, but the most compelling problem other than the complications involved in developing a new drug has been the factor time and capital cost.

Quantitative Structure-Activity Relationship (QSAR) analysis is one of the most effective approaches for optimizing leading compounds and designing new drugs. QSAR can be employed in predicting the bioactivity such as toxicity, mutagenicity, and carcinogenicity based on structural parameters of compounds and appropriate mathematical models. With the rapid development of computer science and theoretical quantum chemical study, it can speedily and precisely obtain the quantum chemical parameters of compounds by computation. Moreover, these parameters, which have definite physical meaning, along with the introduction of the QSAR model can increase the chances of predicting the activities of the object compound and so quantum chemical theory is extensively applied in establishing QSAR models. The aim of this research is to find a new model which predicts the toxicity of chemicals with potent activities able to destroy K562 leukemic cell line using Genetic Algorithm-Multiple Linear Regression (GA-MLR) technique [5-8].

MATERIALS AND METHODS

Data sources

In this study, a data set of one hundred and twelve (112) anticancer compounds collected from the National Cancer Institute (NCI) database. These chemical structures were aligned with their respective bioactive component values on a 2D table after they 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 [9, 10]. 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 [11]. 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 [12, 13]. A distinctive feature of GFA is that it produces a population of models, instead of generating a single model, so do most other statistical methods. The established models were then subjected to internal and external validation and Y-randomization tests in order to institute their predictability and reliability [14].

Geometric 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) [15] 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" [16]. Different physicochemical descriptors were calculated for each molecule presented in Table 1. These descriptors included electronic, spatial, structural, and topological chemical descriptors. These were combined to the set of quantum chemical descriptors obtained from the low energy conformer of the structures 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 predictability 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 [17], [18].

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" [19]. 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} and pLC_{50}) were subjected to multivariate analysis by Genetic Function Approximation (GFA) technique using the material studio software. GFA measures the fitness of a model during the evolution process by calculating the Friedman lack-of-fit (LOF). In Materials Studio, LOF is calculated using the expression:

$$LOF = \frac{SSE}{\left(1 - \frac{C + dp}{m}\right)^2}$$

Where ***SSE*** is the sum of squares of errors, ***C*** is the number of terms in the model, other than the constant term, and is a user-defined smoothing parameter, ***dp*** is the total number of descriptors contained in all model terms (again ignoring the constant term) and ***m*** is the number of samples in the training set [20].

Internal 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 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 $\bar{Y}$ corresponds to the mean observed activity of the training set compounds.

External Model Validation

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 R^2(R^2pred) 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 the 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 found that R^2_{pred} may not be sufficient to be indicated the external predictability of a model since its value is controlled by $\sum\left(Y_{(Test)} - \bar{Y}_{(Training)}\right)^2$.

RESULT AND DISCUSSION

Geometry Optimization and Descriptors Calculation

The observed activities for the various data sets were transformed to obtain a more uniformly distributed data as shown in Table 2. After minimization of the various compounds in the data set 32 descriptors were generated using the Spartan 14 software. These were combined to the 1875 descriptors generated using the paDEL software to give a total of 1907 descriptors.

Table 1: Experimental and Predicted toxicities (pLC_{50}) on different leukemia cell lines obtained with linear models based on GA-MLR technique.

Serial Number (ID)	NSC	K562 (Experimental pLC_{50})	K562 (Predicted pLC_{50})
1	606172	4.0	4.199
2	606173	4.0	4.013
3	643833	4.9	4.361
4	27640	2.6	2.815
5	95678	3.0	3.140
6	264880	2.6	2.592
7	127716	3.4 [b]	1.639
8	102816	2.7	3.065
9	107392	2.8	3.046
10	249910	4.0	4.203
11	629971	4.0	3.508
12	163501	3.0 [b]	3.125
13	406042	4.0 [b]	3.538
14	71851	2.3	2.951
15	132483	4.0	3.844
16	184692	4.0	4.427
17	134033	4.0 [b]	4.062
18	308847	3.6	3.457
19	623017	4.0	4.053
20	355644	4.1	3.946
21	303812	4.0	4.160
22	63878	3.3	2.635
23	167780	3.9 [b]	3.270
24	182986	3.7	3.906
25	139105	3.0	3.047
26	409962	3.4	3.105

27	71261	2.9	2.795
28	337766	4.3 [b]	4.167
29	368390	3.3 [b]	2.706
30	750	3.6 [b]	2.740
31	94600	4.0	4.209
32	295500	4.0	4.172
33	606985	4.0	3.846
34	295501	4.0	3.708
35	606499	4.0 [b]	4.168
36	606497	4.0 [b]	4.118
37	610459	4.0 [b]	4.043
38	610456	4.0 [b]	3.781
39	610457	4.0 [b]	4.957
40	610458	5.0 [b]	4.357
41	176323	4.0	4.082
42	95382	4.0	3.630
43	107124	4.1	4.025
44	100880	3.6	3.579
45	374028	4.0	4.179
46	618939	5.0	4.991
47	79037	3.3	3.279
48	3088	3.1	3.511
49	178248	2.9	3.009
50	338947	2.3 [b]	2.751
51	757	3.2	3.430
52	33410	4.9 [b]	3.677
53	357704	4.6	4.499
54	145668	3.0 [b]	2.707
55	348948	2.6 [b]	3.337
56	82151	4.2	4.247

57	267469	3.9	4.396
58	132313	3.8	3.221
59	126771	3.6	3.437
60	376128	*8.0	8.000
61	123127	4.7	4.417
62	73754	2.6	2.663
63	148958	3.0	3.259
64	364830	4.0	4.226
65	1895	2.0 [b]	2.187
66	329680	2.6	2.585
67	142982	4.2	4.107
68	32065	2.6	2.458
69	118994	2.6	2.423
70	153353	3.3 [b]	2.246
71	330500	4 [b]	4.145
72	249992	3.8	3.790
73	153858	4.0	3.668
74	8806	3.6	3.173
75	269148	4.1	4.322
76	740	3.6	3.412
77	174121	7.0	6.731
78	95441	3.6	3.432
79	26980	4.6	3.807
80	301739	4.9	4.877
81	353451	2.9	3.464
82	354646	5.0	4.962
83	224131	2.0	1.988
84	268242	*4.3	4.712
85	762	3.4	3.321
86	349174	3.6	4.015

87	95466	2.9	2.629
88	344007	3.0	3.180
89	135758	3.0	3.014
90	25154	3.3	3.742
91	56410	3.1	3.747
92	143095	2.3	2.567
93	366140	4.4	4.008
94	51143	2.0	2.180
95	332598	4.0	3.489
96	164011	4.1	3.988
97	172112	3.6	3.637
98	125973	4.6	3.895
99	296934	2.6 [b]	3.727
100	363812	*3.6	3.546
101	361792	4.0	3.649
102	752	3.6	3.868
103	6396	3.0	2.951
104	9706	4.0	3.845
105	352122	3.7	3.739
106	83265	3.9	3.499
107	34462	3.3	3.577
108	49842	5.6	5.458
109	67574	3.2	3.932
110	122819	4.6	4.313
111	141540	3.0	3.585
112	102627	2.0	2.132

Where superscript **letters (b)** represent test sets for the leukemia cell line, and ***** identifies compounds found outside the applicability domain (outliers) of the model, while **NSC** - represents the NCI's internal identification number of the database entry, and is derived from (part of) the acronym of the Cancer Chemotherapy National Service Center (CCNSC).

Feature Selection and Data Division

The created descriptor outcomes were exposed to data pre-treatment where descriptors with constant value and pairs of variables with correlation coefficient greater than 0.7 were removed using the software ***"Data Pre-Treatment GUI 1.2"*.** This was done to be devoid of the model of intercorrelated descriptors. Data pre-treatment resulted in 681 descriptors from 1907 descriptors, thus removing 1226 invariable and highly correlated descriptors. Data division using "Dataset Division GUI 1.2" tool resulted in 90 molecular compounds (covering about 80% of the entire compounds) in the training set and 22 compounds (covering about 20% of the entire compounds) in the test set.

Model Development and Validation

About two hundred and fifty models were generated from the training set by Genetic Function Approximation using the Material Studio Software and the best model based on internal validatory statistical parameters was selected for their toxicity (LC_{50}). The developed model and the description of the molecular descriptors were shown in the equation pLC_{50} below with the model statistics. The predicted values for the training set by the QSAR model was generated by the Material Studio Software, while the predicted test set values was calculated using MSExcel 2013 [21] as reported in Table 2. The results for the model validation of the developed models are given as follows.

$$\begin{aligned} p\mathrm{LC}_{50} = \; & 6.602\,(\mathbf{Secondary\ butyl}) - 1.513\,(\mathbf{E-LUMO}) - 0.892\,(\mathbf{ALogp2}) \\ & + 1.560\,(\mathbf{GATS5p}) - 2.566\,(\mathbf{minHBint7}) + 0.795\,(\mathbf{maxHBint7}) \\ & - 0.539\,(\mathbf{maxHBint8}) + 1.503\left(\mathbf{ETA_{EtaP_L}}\right) + 1.159\,(\mathbf{nF10Ring}) \\ & + 4.269\,(\mathbf{WPSA-3}) - 3.795\,(\mathbf{RDF140u}) + 1.274\,(\mathbf{RDF145m}) + 3.079 \end{aligned}$$

$$N_{train} = 90, \quad R^2_{train} = 0.888,\ adjR^2_{train} = 0.871,\ F_{train} = 50.976, \quad Q^2_{CV} = 0.845,$$

$$N_{test} = 22,\ \text{Outliers} = 03$$

The high calculated Q^2_{LOO} value (0.845) for pLC_{50} proposes a good internal validation. A second validation method was also developed on the basis of an external validation method, here the test set constituting 20% of the data set were subjected to the developed model and the

result were found promising, since its value 0.532 which was higher than the standard value 0.50 for the toxicity model.

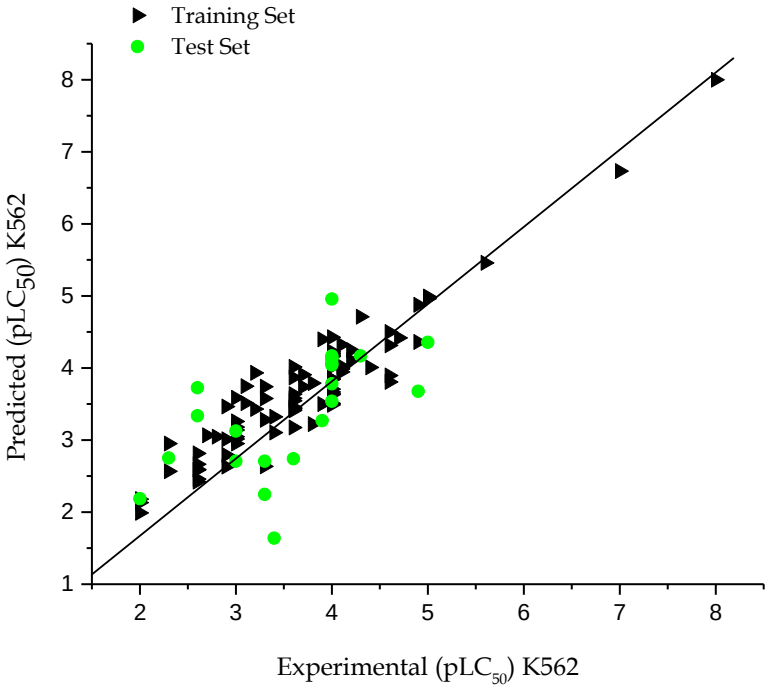

Figure 1: The predicted toxicity values (pLC_{50}) against the experimental values for the training and test sets of the compounds on K562 leukemia cell line.

These values indicate the robustness and stability of the constructed models, as can be seen that the model did not show any proportional and systematic errors, because the propagation of the residuals on both sides of zero is random. The built model was used to predict the test set data, and the prediction results are given in Table 1. The predicted values for pLC_{50} for the compounds in the training and test sets using pLC_{50} equation were plotted against the experimental pLC_{50} values in Figure 1, the calculated values for the pLC_{50} is in good agreement with those of the experimental values. Also, the plot of the standardized residual and leverages values for pLC_{50} is shown in Figure 2.

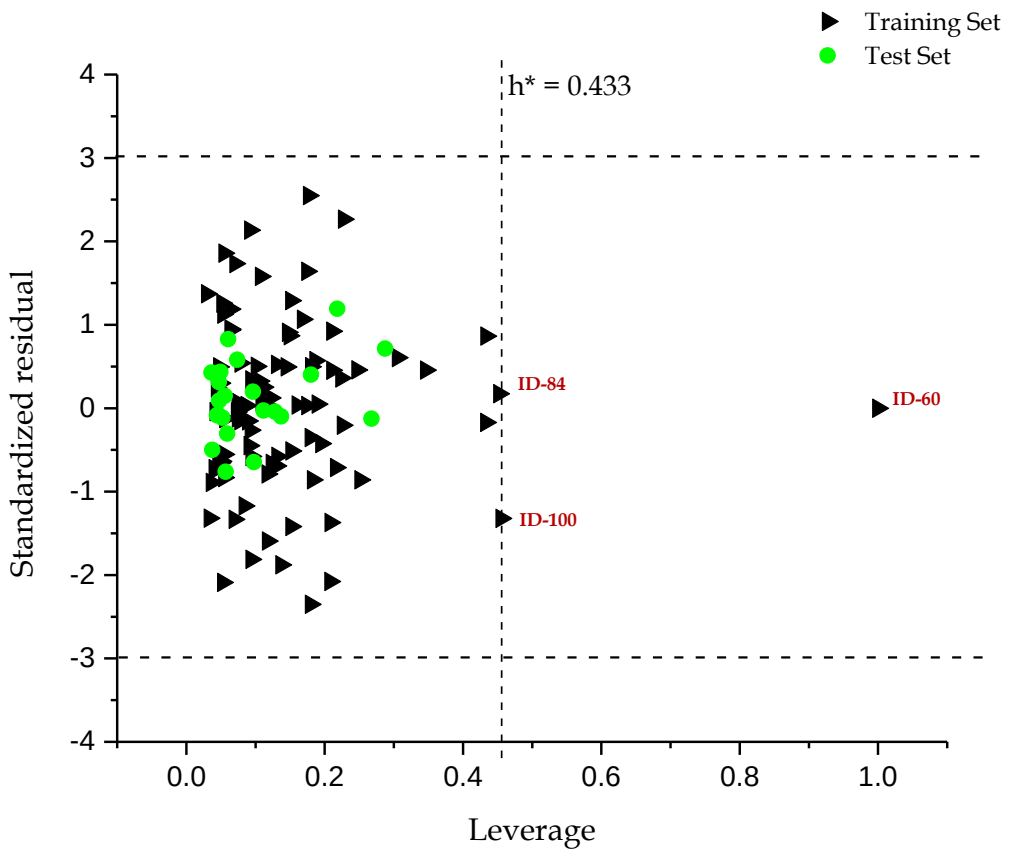

Figure 2: The Williams plot, the plot of the standardized residuals versus the Toxicity (pLC_{50}) leverage value for K-562 dataset.

The Williams plot in Figure 2 shows that the only three compounds were found outside the applicability domain of the molecule, these compounds with ID numbers 60, 84 and 100 were part of the training set. The plot indicates that these compounds structurally different, that is these compounds found outside the threshold value h*, have very few of the chemical descriptors which could be related to those in the model when compared to other compounds within the complete data set.

Table 2: External Validation Result for K-562 cell line.

Model biasness test	Systematic Error Result	Absent
Classical Metrics (for 100% data)	**R^2Test(100% data)**	0.5316
	R0^2Test(100% data)	0.4473
	Q2F1(100% data)	0.4268
	Q2F2(100% data)	0.4255
	Scaled Avg.Rm^2(100% data)	0.4003
	Scaled DeltaRm^2(100% data)	0.0755
	CCC(100% data)	0.7137
Error-based metrics (for 100% data)	**RMSEP(100% data)**	0.5769
	SD(100% data)	0.3636
	SE(100% data)	0.0813
	MAE(100% data)	0.4552

The external validation of the model in Table 2, showing that the mean absolute error (MAE) value is 0.455. Since the value is less than unity, deductions could be made that the predictions are close to the experimental outcomes and thereby supporting the value of the regression coefficient of the predicted test set (0.532), other classical statistical metrics such as Q^2_{F1} and Q^2_{F2} presented in Table 2, supports the result already stated in the discussion.

Interpretation of descriptors

Secondary butyl is a 2D molecular descriptor used in the model, it's defined as the number of secondary butyl group found in potent anticancer compounds. The mean effect reported in Table 3 indicates that the increase presence of this property diminishes the toxicity property of these molecules. Chemical descriptors like E LUMO and ALogp2 defined as energy of lowest occupied molecular orbital and square of AlogP respectively were also used in modelling the pLC50 property of these compounds gave a negative contribution to the model (-1.123 and -0.340).

GATs5p is 2D autocorrelation defined as Geary autocorrelation - lag 5 / weighted by polarizabilities, the descriptor was first stated by Todeschini and Consonni [22], in the book titled Molecular descriptors for chemoinformatics. It is an autocorrelation type descriptor depending on the polarizing ability of the active sites of a chemical drug compound. The descriptor was found to contribute positively to the model which is owned to the value of its

mean effect. The mean effect of GATs5p indicates that its presence decreases toxicity in anticancer drugs.

minHBint7, maxHBint7 and maxHBint8 are 2D Atom type electrotopological state molecular descriptors defined as Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7, Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 7 and Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 8, respectively. Their mean effect was calculated and reported respectively as -1.637, 0.351 and -0.226. The mean effects of minHbint7 and maxHbint8 were negative indicating their direct involvement with the toxicity of the modeled compounds. The value of minHBint7 was most significant of all the three E-state type descriptors related to specific hydrogen bonds in a certain path length.

ETA_EtaP_L is a 2D Extended topochemical atom type descriptor defined as Local index Eta_local relative to molecular size, hence this topochemical descriptor is depend on the molecular size of the molecules. The mean effect given as 1.054 indicates that an increase in the molecular size of this type descriptor will decrease the toxicity of anticancer compounds.

nF10Ring, WPSA-3 and RDF140u which are the final descriptors in the model give the contributions 0.454, 1.657 and -0.705 respectively. They defined as the number of 10-membered fused rings, PPSA-3 multiplied by the total molecular surface area/1000 and the radial distribution function -140/unweighted respectively. WPSA-3 was found to give the highest contribution in the model, its value correlates with that of ETA_EtaP_L significantly, this similarity can be seen from their meanings which is a subject of the molecular size of this active compounds, thereby completing agreeing with the fact that large size molecules have less tendency of being toxic when used as anticancer drugs.

Table 3: Specification of entered descriptors in genetic algorithm multiple regression model of K-562.

Descriptors	**Definition**	**ME**
Secondary butyl	Number of secondary butyl group	0.119
E LUMO	Energy of lowest occupied molecular orbital	-1.123
ALogp2	Square of ALogP	-0.340
GATS5p	Geary autocorrelation - lag 5 / weighted by polarizabilities	1.227
minHBint7	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7	-1.637
maxHBint7	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 7	0.351
maxHBint8	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 8	-0.226
ETA_EtaP_L	Local index Eta_local relative to molecular size	1.054
nF10Ring	Number of 10-membered fused rings	0.454
WPSA-3	PPSA-3 * total molecular surface area / 1000	1.657
RDF145m	Radial distribution function - 145 / weighted by relative mass	0.554
RDF140u	Radial distribution function - 140 / unweighted	-0.705

CONCLUSION

The pLC_{50} for the leukemia cell line K562 was positively modelled for a sequence of anticancer drugs collected from NCI library, using highly interconnected descriptors computed using paDEL software, the statistical parameters of the model satisfy the criteria proposed by Tropsha, Roy and Grammatica for validating QSAR models. A few descriptors such as WPSA-3, minHBint7, GATS5p, E LUMO and ETA_EtaP_L with mean effects of 1.657, -1.637, 1.227, -1.123 and 1.054 respectively, were found to be significantly responsible for the toxicity of the compounds used in the data set.

COMPETING INTERESTS

The authors have declared no conflict of interest.

REFERENCES

1. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS. Rational drug design for anti-cancer chemotherapy: Multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. Bioorganic & Medicinal Chemistry. 2012 Aug;20(15):4848–55. DOI: 10.1016/j.bmc.2012.05.071.

2. Dunnington BD, Schmidt JR. Molecular bonding-based descriptors for surface adsorption and reactivity. Journal of Catalysis. 2015 Apr;324:50–8. DOI: 10.1016/j.jcat.2015.01.017.

3. Andrada MF, Vega-Hissi EG, Estrada MR, Garro Martinez JC. 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 and Intelligent Laboratory Systems. 2015 Apr;143:122–9. DOI: 10.1016/j.chemolab.2015.03.001.

4. Alanazi AM, Abdel-Aziz AA-M, Al-Suwaidan IA, Abdel-Hamide SG, Shawer TZ, El-Azab AS. Design, synthesis and biological evaluation of some novel substituted quinazolines as antitumor agents. European Journal of Medicinal Chemistry. 2014 May;79:446–54. DOI: 10.1016/j.ejmech.2014.04.029.

5. Gagic Z, Nikolic K, Ivkovic B, Filipic S, Agbaba D. QSAR studies and design of new analogs of vitamin E with enhanced antiproliferative activity on MCF-7 breast cancer cells. Journal of the Taiwan Institute of Chemical Engineers. 2016 Feb;59:33–44. DOI: 10.1016/j.jtice.2015.07.019.

6. Chen B, Zhang T, Bond T, Gan Y. Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: A review of methods and resources. Journal of Hazardous Materials. 2015 Dec;299:260–79. DOI: 10.1016/j.jhazmat.2015.06.054.

7. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS. Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the in silico discovery of anti-breast cancer agents. European Journal of Pharmaceutical Sciences. 2012 Aug;47(1):273–9. DOI: 10.1016/j.ejps.2012.04.012.

8. Zhao L, Xiang Y, Song J, Zhang Z. A novel two-step QSAR modeling work flow to predict selectivity and activity of HDAC inhibitors. Bioorganic & Medicinal Chemistry Letters. 2013 Feb;23(4):929–33. DOI: 10.1016/j.bmcl.2012.12.067.

9. Benarous N, Cherouana A, Aubert E, Durand P, Dahaoui S. 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 Feb;1105:186–93. DOI: 10.1016/j.molstruc.2015.10.037.

10. Bauernschmitt R, Ahlrichs R. Treatment of electronic excitations within the adiabatic approximation of time dependent density functional theory. Chemical Physics Letters. 1996 Jul;256(4–5):454–64. DOI: 10.1016/0009-2614(96)00440-X.

11. Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics. 1969 Feb;11(1):137–48. DOI: 10.1080/00401706.1969.10490666.

12. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002 Apr;6(2):182–97. DOI: 10.1109/4235.996017.

13. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. Journal of Chemometrics. 1992 Sep;6(5):267–81. DOI: 10.1002/cem.1180060506.

14. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics. 2010 Jul 6;29(6–7):476–88. DOI: 10.1002/minf.201000061.

15. Hehre WJ, Huang WW. Chemistry with computation: an introduction to SPARTAN. Irvine: Wavefunction, Inc.; 1995. ISBN: 9780964349520.

16. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 2011 May;32(7):1466–74. DOI: 10.1002/jcc.21707.

17. Panagos P, Meusburger K, Ballabio C, Borrelli P, Alewell C. Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Science of The Total Environment. 2014 May;479–480:189–200. DOI: 10.1016/j.scitotenv.2014.02.010.

18. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems. 2015 Jul;145:22–9. DOI: 10.1016/j.chemolab.2015.04.013.

19. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar 28;483(7391):603–307. DOI: 10.1038/nature11003.

20. Abdel-Atty MM, Farag NA, Kassab SE, Serya RAT, Abouzid KAM. Design, synthesis, 3D pharmacophore, QSAR, and docking studies of carboxylic acid derivatives as Histone Deacetylase inhibitors and cytotoxic agents. Bioorganic Chemistry. 2014 Dec;57:65–82. DOI: 10.1016/j.bioorg.2014.08.006.

21. Carlberg C. Statistical analysis: microsoft excel 2013. 1st edition. Indianapolis, IN: Que Pub; 2014. ISBN: 9780789753113.

22. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. Weinheim: Wiley-VCH; 2009. (Methods and principles in medicinal chemistry). ISBN: 9783527318520.

Türkçe Öz ve Anahtar Kelimeler

Genetik Algoritma-Çoklu Lineer Regresyon Kullanarak K562 Kanser Hücre Dizisine Karşı Bazı Aktif Bileşiklerin Zehirlilik Modellemesi

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

Öz: Bu araştırmada antikanser bileşiklerinin K562 hücre dizisi üzerindeki zehirliliğinin modellemesi araştırılacaktır, veri serisini oluşturan 112 bileşik, modelin sırasıyla geliştirilmesi ve doğrulanması için eğitim ve test setleri olarak ayrılmıştır. İç ve dış doğrulama parametresi olan R^2, eğitim ve test serisi için 0,845 ve 0,5316 olarak tespit edilmiştir, bu da modelin sağlamlığını ve bileşiklerin zehirliliğini tahmin etme yeteneğini belirlemek için kullanılmıştır. Modelin ortalama etkisinin %50 kadarından WPSA-3 ve minHBint7 moleküler tarifçi sorumludur.

Anahtar kelimeler: QSAR; modelleme; dış doğrulama; moleküler tarifiçiler; genetik algoritma.

Sunulma: 05 Ekim 2016. **Düzeltme:** 17 Kasım 2016. **Kabul:** 21 Kasım 2016.