Turkish Computational and Theoretical Chemistry

Turkish Comp Theo Chem (TC&TC)

Volume(Issue): 3(1) – Year: 2019 – Pages: 38-46

e-ISSN: 2602-3237

https://doi.org/10.33435/tcandtc.458664

Received: 10.09.2018 ***Accepted***: 11.12.2018 *Research Article*

Multivariant QSAR Model for Some Potent Compounds as Potential Anti-Tumor Inhibitors: A Computational Approach

Shola Elijah ADENIJI[1], Sani UBA, Adamu UZAIRU

Department of Chemistry, Ahmadu Bello University, Zaria-Nigeria

Abstract: A computational approach was employed to develop multivariate QSAR model to correlate the chemical structures of the ciprofloxacin analogues with their observed activities using a theoretical approach. Genetic Function Algorithm (GFA) and Multiple Linear Regression Analysis (MLRA) were used to select the descriptors and to generate the correlation QSAR models that relate the activity values against tumor with the molecular structures of the active molecules. The models were validated and the best model selected has squared correlation coefficient (R2) of 0.990531, adjusted squared correlation coefficient (Radj) of 0.95962 and Leave one out (LOO) cross validation coefficient (Q_cv^2) value of 0.942963. The external validation set used for confirming the predictive power of the model has its R2pred of 0.8486. Stability and robustness of the model obtained by the validation test indicate that the model can be used to design and synthesis other ciprofloxacin derivatives with improved anti-tumor activity.

Keywords: *Ciprofloxacin, Descriptors, Genetic Function Algorithm, tumor, QSAR*

1. Introduction

Prostate cancer as one of the leading tumor develops when abnormal cells in the prostate gland start to grow more rapidly than normal cells, and in an uncontrolled way. Prostate Cancer has been reported as a major tumor in men with significant incidence and morbidity [1]. It diagnosed primarily in older men, with a majority being over age 65, although men in their 30s and 40s have been diagnosed with the disease. Its incidence and prevalence in black men is in multiples of those from other races in several studies [2]. The reason for this is not yet clear and an explanation for the disparity may lie in studies involving black men from different populations to see if there is an enhancing factor associated with the racial origins of these men.

Ciprofloxacin (CP), an antibiotic has been shown to have anti-proliferative and apoptotic activities in several cancer cell lines. Moreover, several reports have highlighted the interest of increasing the lipophilicity to improve the antitumor efficacy.

Synthesis of novel compounds are developed using a trial and error approach, which is time consuming and expensive. The application of Quantitative Structure Activity Relationship (QSAR) technique to this problem has potential to minimize effort and time required to discover new compounds or to improve current ones in terms of their efficiency. QSAR establishes the mathematical relationship between physical, chemical, biological or environmental activities of interest and measurable or computable parameters such as physicochemical, topological, stereo chemical or electronic indices called molecular descriptors [3]. The aim of this research was to

[1] Corresponding Author

e-mail: shola4343@gmail.com

develop various QSAR models for predicting the activity of ciprofloxacin derivatives against tumor.

2. Materials and Method

2.1. Data Collection

Data set of ciprofloxacin derivatives as potential anti-tumor that were used in this study were obtained from the literature [4].

2.2. Biological Activities (pIC_{50})

The Biological activities of ciprofloxacin derivatives against tumor measured in IC_{50} (μM) were converted to logarithm unit (pIC_{50}) using the equation (1) below in order to increase the linearity activities values and approach normal distribution. The observed structures and the biological activities of these compounds were presented in Figure 1 and Table 1.

$$pIC_{50} = -\log (IC_{50}) \quad (1)$$

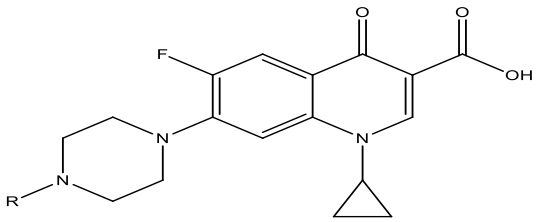

Figure 1. General structure of ciprofloxacin derivatives

Table 1. Molecular structure, Experimental, Predicted and Residual values of ciprofloxacin derivatives as potent anti-tumor

S/N	R	Activity IC_{50} (μM)	Experimental Activity (pIC_{50})	Predicted activity	Residual
1 [a]	H	143	3.844664	3.732084	0.11258
2	$COCH_2Cl$	8	5.09691	5.1541	-0.05719
3	$C(O)OC(CH_3)_3$	26	4.585027	4.681317	-0.09629
4	$COCH_2OCOCH_3$	176	3.754487	3.763877	-0.00939
5	$COCH_2OCO(CH_2)2CH_3$	715	3.145694	3.137145	0.008549
6 [a]	$COCH_2OCO(CH_2)4CH_3$	14	4.853872	4.917432	-0.06356
7	$COCH_2OCO(CH_2)6CH_3$	23	4.638272	4.592276	0.045996
8 [a]	$COCH_3$	680	3.167491	3.12956	0.037931
9 [a]	$COCH_2\ CH_3$	352	3.453457	3.350235	0.103222
10	$CO(CH_2)_2CH_3$	85	4.070581	3.957798	0.112783
11	$CO(CH_2)_3\ CH_3$	73	4.136677	4.215267	-0.07859
12	$COC(CH_3)_3$	246	3.609065	3.428617	0.180448
13	$CO(CH_2)_5CH_3$	779	3.108463	3.063929	0.044534
14	$CO(CH_2)_7CH_3$	7	5.154902	5.304892	-0.14999
15	$CO(CH_2)_8CH_3$	4	5.39794	5.357801	0.040139
16	$CO(CH_2)_{10}CH_3$	4	5.39794	5.360478	0.037462
17	$CO(CH_2)_{12}CH_3$	94	4.026872	4.062242	-0.03537
18 [a]	$CO(CH_2)_{14}CH_3$	114	3.943095	3.239815	0,70328
19	$COCH_2C_6H_5$	243	3.614394	3.58375	0.030644
20 [a]	$COCH_2OH$	433	3.363512	3.122104	0.241408

Where superscript a represent the test set

2.3. Optimization

The 2D structures of the compounds presented in the Table 1 were drawn utilizing chemdraw programming [5]. The spatial conformations of the compounds were exported from 2D structure to 3D format using the Spartan 14 V1.1.4 Wave Function programming package. All 3D structures were geometrically optimized by minimizing energy. The chemical structures were initially minimized by Molecular Mechanics Force Field (MMFF) count to remove strain energy before subjecting it to quantum chemical estimations. Density Functional Theory (DFT) method was later employed by utilizing the Becke's three parameter exchange functional (B3) hybrid with Lee, Yang and Parr correlation functional (LYP) which is termed (B3LYP) hybrid functional for complete geometric optimization of the structures [6,7]. The Spartan files of all the optimized molecules were then saved in SD file format, which is the

recommended input format in PaDEL-Descriptor software V2.20 [8].

2.4. Molecular Descriptor Calculation

Molecular descriptors are mathematical values that describe the properties of a molecule. Descriptors calculation for all the 20 molecules of ciprofloxacin derivatives were calculated using PaDEL-Descriptor software V2.20. A total of 1876 molecular descriptors were calculated.

2.5. Normalization and Data Pretreatment

The descriptors' value were normalized using Equation 2 in order to give each variable the same opportunity at the onset to influence the model [9].

$$X = \frac{X_1 - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Where Xi is the value of each descriptor for a given molecule, Xmax and Xmin are the maximum and minimum value for each column of descriptors X. The normalized data were subjected to pretreatment using Data Pretreatment software obtained from Drug Theoretical and Cheminformatics Laboratory (DTC Lab) in order to remove noise and redundant data [8].

2.6. Data Division

In order to obtain validated QSAR models the dataset was divided into training and test sets using Data Division software obtained from Drug Theoretical and Cheminformatics Laboratory (DTC Lab) by employing Kennard and Stone's algorithm . This algorithm has been applied with great success in many recent QSAR studies and has been highlighted as one of the best ways to build training and test sets [10–14]. In this algorithm, two compounds with the largest Euclidean distance apart were initially selected for the training set. The remaining compounds for the training set were selected by maximizing the minimum distance between these two compounds and the rest of the compounds in the dataset. This process continues until the desired number of compounds needed for the training set have been selected then, the remaining compounds in the dataset would be used as the test set.

The algorithm employs Euclidean distance ED_X (p, q), between the x vectors of each pair (p, q) of samples to ensure a uniform distribution of such a subset along the x data space

$$ED_X(p, q) = \sqrt{\sum_{j=1}^{N} [x_p(j) - x_q(j)]} \; 2p,q \in [l, m] \quad (3)$$

N is the number variables in x, and m is the number of samples while $x_p(j)$ and $x_q(j)$ are the *j*th variable for samples p and q respectively.

The training set was used to generate the model, while the test set were used for the external validation of the model.

2.7. Data Division

Validation of the model was carried out using Material studio software version 8 using Genetic Function Approximation (GFA) method [15]. The models were estimated using the LOF, which was measured using a slight variation of the original Friedman formula, so that the best fitness score can be received. In materials studio version 8, Lack of fit (LOF) is measured using a slight variation of the original Friedman formula. The revised formula is:

$$LOF = \frac{SEE}{\left(1 - \frac{C + d \times p}{M}\right)^2} \quad (4)$$

where c is the number of terms in the model, other than the constant term, d is a user-defined smoothing parameter, p is the total number of descriptors contained in the model and M is the number of data in the training set. **SEE** is the Standard Error of Estimation which is equivalent to the models standard deviation. It's a measure of model quality and a model is said to be a better model if it has low SEE value. SEE is defined by equation below;

$$SEE = \sqrt{\frac{\left(Y_{exp} - Y_{pred}\right)^2}{N - P - 1}} \quad (5)$$

The square of the correlation coefficient (R^2) describes the fraction of the total variation attributed to the model. The closer the value of R^2 is to 1.0, the better the regression equation explains the Y variable. R^2 is the most commonly used internal validation indicator and is expressed as follows:

$$R^2 = 1 - \left[\frac{\Sigma\left(Y_{exp} - Y_{pred}\right)^2}{\Sigma\left(Y_{exp} - \overline{Y}_{training}\right)^2}\right] \quad (6)$$

where Y_{exp}, Y_{pred} and $\overline{Y}_{training}$ are the experimental activity, the predicted activity and the mean experimental activity of the samples in the training set, respectively.

R^2 value varies directly with the increase in number of repressors i.e. descriptors, thus, R^2 cannot be a useful measure for the stability of model. Therefore, R^2 is adjusted for the number of explanatory variables in the model. The adjusted R^2 is defined as:

$$R^2_{adj} = \frac{R^2 - P\,(n-1)}{n - p + 1} \quad (7)$$

where p is the number of independent variables in the model. The capability of the QSAR equation to predict bioactivity of new compounds was determined using the leave-one-out cross validation method. The cross-validation regression coefficient (Q^2_{cv}) was calculated with the equation below:

$$Q^2_{cv} = 1 - \left[\frac{\Sigma\left(Y_{pred} - Y_{exp}\right)^2}{\Sigma\left(Y_{exp} - \overline{Y}_{training}\right)^2}\right] \quad (8)$$

The coefficient of determination for the test set R^2_{test} was calculated with the equation below;

$$R^2_{test} = 1 - \frac{\Sigma\left(Ypred_{test} - Yexp_{test}\right)^2}{\Sigma\left(Ypred_{test} - \overline{Y}_{training}\right)^2} \quad (9)$$

2.8. Y-Randomization Test

To guarantee the created QSAR model is strong and not inferred by chance, the Y-randomization test was performed on the training set data as suggested by [16]. Random MLR models are generated by randomly shuffling the dependent variable (activity data) while keeping the independent variables (descriptors) unaltered. The new QSAR models are expected to have significantly low R^2 and Q^2 values for several trials, which confirm that the developed QSAR models are robust. Another parameter, cR^2_p is also calculated which should be more than 0.5 for passing this test.

$$cR^2_p = R \times [R^2 - (R_r)^2]^2 \quad (10)$$

where cR^2_p is the coefficient of determination for Y-randomization, R; coefficient of determination for Y-randomization and Rr; average 'R' of random models.

2.9. Quality Assurance of The Model

The fitting ability, stability, reliability and predictive ability of the developed models were evaluated by internal and external validation parameters. The validation parameters were compared with the minimum recommended value for a generally acceptable QSAR model [17] showed in Table 2.

Table 2. Minimum recommended value of Validation Parameters for a generally acceptable QSAR model

Symbol	Name	Value
R^2	Coefficient of determination	≥ 0.6
P (95%)	Confidence interval at 95% confidence level	< 0.05
Q^2_{cv}	Cross validation coefficient	> 0.5
R^2 - Q^2_{cv}	Difference between R^2 and Q^2_{cv}	≤ 0.3
$N_{ext.\ test\ set}$	Minimum number of external test set	≥ 5
cR^2_p	Coefficient of determination for Y-randomization	> 0.5

3. Results

Table 3. Validation parameters from material studio

S/N	Validation Parameters	Model 1	Model 2	Model 3	Model 4
1	Friedman LOF	0.287447	0.29417	0.319241	0.36543
2	R-squared	0.990531	0.948212	0.875503	0.82954
3	Adjusted R-squared	0.95962	0.958676	0.955154	0.91245
4	Cross validated R-squared	0.942963	0.935828	0.934816	0.87353
56	Significant Regression	Yes	Yes	Yes	Yes
7	Significance-of-regression F-value	103.980981	101.528362	93.293244	91.3344
8	Critical SOR F-value (95%)	3.871034	3.871034	3.871034	3.871034
9	Replicate points	0	0	0	0
10	Computed experimental error	0	0	0	0
11	Lack-of-fit points	10	10	10	0
12	Min expt. error for non-significant LOF (95%)	0.186643	0.208814	0.266695	0.31900

Table 4. List of some descriptors used in the QSAR optimization model

S/NO	Descriptors symbols	Name of descriptor(s)	Class
1	AATSC6m	Average centered Broto-Moreau autocorrelation - lag 6 / weighted by mass	2D
2	MDEC-22	Molecular distance edge between all secondary carbons	2D
3	L3v	3rd component size directional WHIM index / weighted by relative van der Waals volumes	3D

Table 5. Pearson's correlation matrix and statistics for descriptor used in the QSAR optimization model

	Inter-correlation			**Statistics**	
Descriptors	AATSC6m	MDEC-22	L3v	VIF	P- value
AATSC6m	1			2.56436	3.34E-05
MDEC-22	-0.15654	1		1.84743	4.23E-04
L3v	-0.19444	0.45585	1	2.34556	5.34E-07

3.1. "Y-Randomization Parameter Test

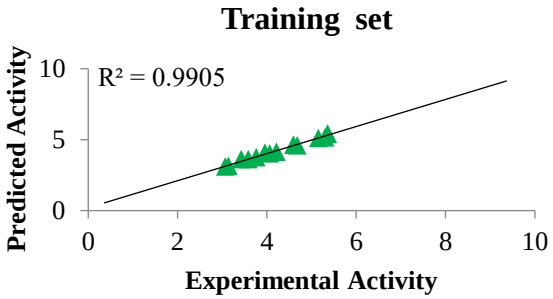

Figure 2. Plot of predicted activity against experimental activity of training set.

Table 6. Y- Randomization Parameters test

Model	R	R^2	Q^2
Original	0.965475	0.932142	0.831909
Random 1	0.674003	0.45428	-0.31323
Random 2	0.61843	0.382455	-0.50841
Random 3	0.311542	0.097058	-1.37797
Random 4	0.632995	0.400683	-0.27203
Random 5	0.665103	0.442362	-0.76461
Random 6	0.385191	0.148372	-1.09687
Random 7	0.583435	0.340396	-0.68669
Random 8	0.446102	0.199007	-1.00243
Random 9	0.413199	0.170734	-0.91905
Random 10	0.788129	0.621147	0.008176
Random Models Parameters			
Average r :	0.551813		
Average r^2 :	0.325649		
Average Q^2 :	-0.69331		
cRp^2 :	0.764888		

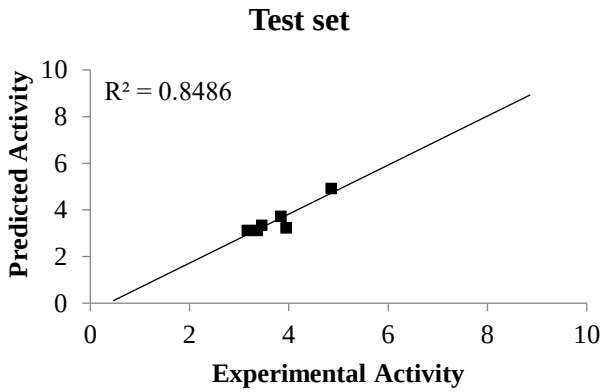

Figure 3. Plot of predicted activity against experimental activity of test set

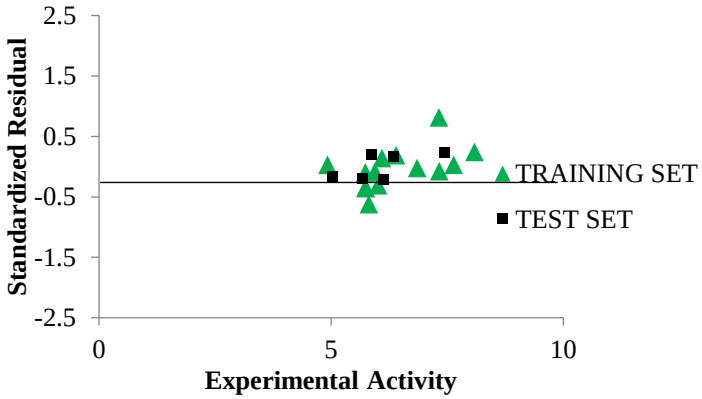

Figure 4. Plot of standardized residual activity versus experimental activity.

4. Discussion

A QSAR examination was performed to investigate the structure activity relationship of 20 compounds as potent anti-tumor. The nature of models in a QSAR study is expressed by its fitting and forecast capacity. In order to assemble a decent QSAR model for anti-tumor with good predictive power for the selected test set. Kennard-Stone algorithm was used to divide the dataset of 20 compounds into a training set of 14 compounds which was used to developed the model and a test set of 6 compounds which was applied to assess the predictive ability built model.

Experimental and Predicted activity for ciprofloxacin derivatives as a potent anti-tumor and the residual values were presented in Table 1. The low residual value between Experimental and Predicted activity indicates that the model is of high predictability.

The Genetic Algorithm- Multi Linear Regression (GA–MLR) investigation led to the selection of three descriptors which were used to assemble a linear model for calculating predictive activity on tumor. Four QSAR models were built using Genetic Function Algorithm (GFA), but due to the statistical significance, model 1 was selected, reported and its parameters were as well calculated.

Model 1

pIC50 = 0.295441891 * AATSC6m + 0.193350923 * MDEC-22 - 1.938081244 * L3v + 7.423458362

Model 2

pIC50 = 0.279119413 * AATSC6m + 0.456910158 * nssCH2 - 1.455230092 * L3v + 7.681216809

Model 3

pIC50 = 0.002899338 * ATSC6v + 0.472513415 * nssCH2 - 1.368491011 * nssCH2 + 8.284970195

Model 4

pIC50 = 0.277548931 * AATSC6m + 0.484912043 * nssCH2 - 1.936444918 * L3v + 6.909123060

External validation and internal validation parameters to confirm that the built QSAR models are stable and robust were reported in Table 3. These parameters were in agreement with the threshold value reported in Table 2 which actually confirmed the robustness and stability of the model.

The name and symbol of the descriptors used in the QSAR optimization model was reported in Table 4. The presence of the 2D and 3D descriptors in the model suggests that these types of descriptors are able to characterize better anti-tumor activities of the compounds. Pearson's correlation matrix and statistics of the three descriptors employed in the QSAR Model were reported in Table 5 which shows clearly that the correlation coefficients between each pair of descriptors is very low thus, it can be inferred that there exist no significant inter-correlation among the descriptors used in building the model [18]. The estimated Variance Inflation Factor (VIF) values for all the descriptors were less than 4 which imply that the Model generated was statistically significant and the descriptors were orthogonal. The p-value is a probability that measures the evidence against the null hypothesis. Lower probabilities provide stronger evidence against the null hypothesis. The null hypothesis implies that there is no association between the descriptors and the activities of the molecules. The P-values of all the descriptors in the model at 95% confidence level shown in Table 5 are less than 0.05. This implies that the alternative hypothesis is accepted. Hence there is a relationship between the descriptors used in the model and the activities molecules which take preference over the null hypothesis[18].

Y- Randomization parameter test were reported in Table 6. The low R^2 and Q^2 values for several trials confirm that the developed QSAR model is robust. While the cR_p^2 value greater than 0.5 affirms that the created model is powerful and not inferred by chance.

Plot of predicted activity against experimental activity of training and test set were shown in Figure 2 and Figure 3 respectively. The R^2 value of 0.9905 for training set and R^2 value of 0.8486 for test set recorded in this study was in agreement with GFA derived R^2 value reported in Table 2. This confirms the reliability of the model. Plot of Standardized residual versus experimental activity shown in Figure 4 indicates that there was no systemic error in model development as the spread

of residuals was pragmatic on both sides of zero [19].

5. Conclusion

This work addresses the Quantitative structure activity relationship (QSAR) between ciprofloxacin derivatives and their (pIC_{50}) against tumor. Results from the optimal model showed that the pIC_{50} of the studied molecules against tumor was affected by (AATSC6m, MDEC-22 and L3v) descriptors. The robustness and applicability of QSAR equation has been established by internal and external validation techniques. Stability and robustness of the model obtained by the validation test indicate that the model can be used to design other ciprofloxacin derivatives with improved anti-tumor activity.

References

[1] N.B. Delongchamps, A. Singh, G.P. Haas, Epidemiology of prostate cancer in Africa: another step in the understanding of the disease?, Current Problems in Cancer. 31 (2007) 226–236.

[2] F.T. Odedina, J.O. Ogunbiyi, F.A. Ukoli, Roots of prostate cancer in African-American men., Journal of the National Medical Association. 98 (2006) 539–548.

[3] A. Rathod, Antifungal and Antibacterial activities of Imidazolylpyrimidines derivatives and their QSAR Studies under Conventional and Microwave-assisted, Int J PharmTech Res. 3 (2011) 1942–1951.

[4] J. Azéma, B. Guidetti, J. Dewelle, B. Le Calve, T. Mijatovic, A. Korolyov, J. Vaysse, M. Malet-Martino, R. Martino, R. Kiss, 7-((4-Substituted) piperazin-1-yl) derivatives of ciprofloxacin: synthesis and in vitro biological evaluation as potential antitumor agents, Bioorganic & Medicinal Chemistry. 17 (2009) 5396–5407.

[5] Z. Li, H. Wan, Y. Shi, P. Ouyang, Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and ChemSketch, Journal of Chemical Information and Computer Sciences. 44 (2004) 1886–1890.

[6] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Physical Review B. 37 (1988) 785.

[7] A.D. Becke, Becke's three parameter hybrid method using the LYP correlation functional, J. Chem. Phys. 98 (1993) 5648–5652.

[8] S.E. Adeniji, S. Uba, A. Uzairu, In Silico Study For Investigating and Predicting the activities of 1, 2, 4-triazole derivaties as potent anti-tubercular agents, The Journal of Engineering and Exact Sciences. 4 (2018) 246–254.

[9] P. Singh, Quantitative Structure-Activity Relationship Study of Substituted-[1, 2, 4] Oxadiazoles as S1P1 Agonists, Journal of Current Chemical and Pharmaceutical Sciences. 3 (2013) 334–345.

[10] G. Melagraki, A. Afantitis, K. Makridima, H. Sarimveis, O. Igglessi-Markopoulou, Prediction of toxicity using a novel RBF neural network training methodology, Journal of Molecular Modeling. 12 (2006) 297–305.

[11] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes, Bioorganic & Medicinal Chemistry. 14 (2006) 6686–6694.

[12] A.K. Chakraborti, B. Gopalakrishnan, M.E. Sobhia, A. Malde, 3D-QSAR studies of indole derivatives as phosphodiesterase IV inhibitors, European Journal of Medicinal Chemistry. 38 (2003) 975–982.

[13] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Artificial neural networks in classification of NIR spectral data: design of the training set, Chemometrics and Intelligent Laboratory Systems. 33 (1996) 35–46.

[14] K.F. Khaled, Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model, Corrosion Science. 53 (2011) 3457–3465.

[15] S.E. Adeniji, D.E. Arthur, A. Oluwaseye, Computational modeling of 4-Phenoxynicotinamide and 4-Phenoxypyrimidine-5-carboxamide derivatives as potent anti-diabetic agent

against TGR5 receptor, Journal of King Saud University-Science. (2018).

[16] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, Molecular Informatics. 22 (2003) 69–77.

[17] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C.P. Varghese, R.K. Agrawal, Validation of QSAR models-strategies and importance, International Journal of Drug Design & Discovery. 3 (2011) 511–519.

[18] S.E. Adeniji, S. Uba, A. Uzairu, QSAR Modeling and Molecular Docking Analysis of Some Active Compounds against Mycobacterium tuberculosis Receptor (Mtb CYP121), Journal of Pathogens. (2018) 1–24.

[19] M. Jalali-Heravi, A. Kyani, Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach, Journal of Chemical Information and Computer Sciences. 44 (2004) 1328–1335.