ORICPublications
www.oricpub.com

Online available since 2017/ December /29 at www.oricpub.com

Journal of Chemistry and Materials Research
Vol. 6 (1), 2016, 3–13

ISSN: 2381-3628
JCMR
Journal of Chemistry and Materials Research
www.oricpub.com/jcmr

Original Research

Quantitative Structure-Activity Relationship Analysis of the Anti-tyrosine Activity of Some Tetraketone and Benzyl-benzoate Derivatives Based on Genetic Algorithm-Multiple Linear Regression

Emmanuel Israel Edache*, David Ebuka Arthur and Usman Abdulfatai

Department of Chemistry, Ahmadu Bello University, Zaria-Nigeria

Received 02 October 2016; received in revised 19 October 2016; accepted 10 January 2017

Abstract

A QSAR study is performed on the seriestetraketone and benzyl benzoate derivatives in order to analyze the physicochemical requirements of tyrosinase inhibitors and to provide structural insight into the binding mode of the molecules to the enzyme. All the derivatives in the series were sketched using ChemDraw ultra v12.0.2 module of ChemOffice 2010 and the sketched structures were consequently used for the calculation of molecular descriptors available in QSAR software Spartan'14 and PaDEL-Descriptors software. Quantum, constitutional and topological descriptors for all molecules were calculated using Spartan'14 v1.1.2, 2013 and PaDEL-Descriptors software v2.20, 2011 and correlation between the biological activity and molecular descriptors was found through genetic function approximation adopted by statistical program material studio v7.0. The generated QSAR models revealed that ATS0s, AATS6p, ATSC1i, SpMAD_Dzv and VR1_Dze descriptors have good correlation to the anti-tyrosinase activity. The results obtained by regression analysis indicated that ATSC1i and SpMAD_Dzv is negatively contributing to inhibitory activity thus; enhancement of anti-tyrosinase activity can be achieved by decreasing the respective descriptors. Positive contribution of ATS0s, AATS6p, and VR1_Dze specifies that increase of Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities, Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities and Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities will impart positive influence on the activity.

Keywords: Tyrosinase; Tetraketone and benzyl benzoate; Semi-empirical (PM6); GFA-MLR; QSAR; Applicability Domain

1. Introduction

Tyrosinase (Syn. Polyphenol Oxidase, PPO, EC 1.14.18.1) is an enzyme that catalyses the oxidation of phenols. It is also known as monophenol mono oxygenase. It is a copper contain-ing enzyme present in animal tissues, higher plants and fungi that catalyses the production of melanin [1,2]. Production of mel-anin causes many kinds of skin diseases, such as hyperpigment spots on the face and freckles [3]. Tyrosinase catalyses both the hydroxylation of monophenols to o-di-phenols (monophenelase or cresolase activity) and the oxidation of o-di-phenols to o-quinonesboth using molecular oxygen followed by a series of nonenzymatics steps resulting in the formation of melanin which plays a crucial protective role against skin photocarcinogenesis [4-6]. Tyrosinase may involve in neuromelanin formation in human brain and contribute to neurodegeneration associated with parkinson's disease [6,8]. In fungi, the role of mel-anin is correlated with the differentiation of reproductive organ and spore formation, virulence of pathogenic fungi, and tissue protection after injury. In addition, it causes undesired enzym-atic browning such as injured cut fruits and vegetables which leads to significant decrease in nutritional values [7]. As tyrosinase inhibitors have an increasing importance due to enormous application prospects in recent periods, the various tyrosinase inhibitors are extracted from natural sources and synthesized. Among which some are applicable to pharmaceutical and cosmetic fields [8]. Tyrosinase inhibitors are useful for the treatment of some dermatological disorders associated with melanin hyperpigmentation, wound healing, parasite encapsul-ation and also important in cosmetics for whitening and depigmentation after sunburn.

Lead optimization is a vital component of the drug discovery process in which a chemical showing promise is modified to greatly improve its usefulness as a drug. Computa-tional methods like quantitative structure activity relationships (QSAR) can facilitate this process by elucidating

* Corresponding author. Tel.: +2348066776802.
E-mail address: edacheson2004@gmail.com ***(E.I. Edache)***.

the chemical characteristics that are favorable and unfavorable through statistical analysis of a series of chemicals [9,10]. QSAR methods derive correlations between the properties/ descript-ors of molecules and their biological activities (e.g., inhibition constants or binding affinities). Since the advent of Free Wilson and Hansch analysis, numerous methods have been published in the literature for structure-activity relation-ships modeling [11]. It is a meaningful correlation (model) between a set of independent variables (chemical descriptors) calculated from chemical graphs, and a dependent variable such as binding affinity, log P, or the pKa value whose value one wishes to predict for the compound of interest [12].

The aim of this paper is to find a correlation between molecular and electronic structures of 37 investigated tyrosinase inhibitors (Table 1) which were found to have tyrosinase activity through inhibiting tyrosinase reductase as their inhibition efficiency IC50 was reported [4,6]. Molecular orbital calculations were performed looking for good theoretical parameters to characterize the inhibition property of inhibitors which will be helpful to gain insight into the mechanism of inhibition.

2. Experimental

2.1. Materials

The materials used in this study include; DELL INSPIRON computer system (Intel Pentium), T4500 2.30 GHz processor Dual-core, 3GB ram size on Microsoft windows 10 operating system, Spartan 14 version 1.1.2, Chem Draw ultra version 12.0.1, PaDEL descriptor tool kit version 2.20 and Microsoft office Excel 2013 statistical software, Material Studio (modeling and simulation software) version 7.0, DTC_Euclidean program version 1.0

2.2. Methods

The data set tetraketone and benzyl benzoate derivatives used in this study was taken from the work of [4,6] and is shown in Table 1. This set contains the values of the anti-tyrosinase inhibition potency compounds. The data set was divided into two groups, a training set consisted of 25 compounds and a test set with 12 compounds. The training and test sets were used for the constructing of the models and to evaluate the predictive power of the generated models, respectively. The inhibitory activities in logarithmic scale (pIC50 = log 1/IC50) fall in the range of -0.314 to 2.233, with a mean value of 0.0428.

2.2.1. Molecular Modeling and Generation of Molecular Descriptors

The dual core personal computer equipped with the operating system Windows ten (10) was used for making calculations of this work. Structure of all the compounds was drawn using Chem Draw Ultra module of the program and transferred to Spartan'14 (2013) version 1.1.2 [13] module to create the three-dimensional (3D) structure. These structures were then subjected to energy minimization using molecular mechanics (MMFF). Energy minimized molecules were subjected to optimization via parameterization method (PM6) [14,15]. These methods have become popular in recent years because they can reach similar precision to other methods in less time and less cost from the computational point of view. The geometry optimization of the lowest energy structure was carried out without any symmetry constraints were also transferred to PaDEL-Descriptor [16] version 2.20 and were subjected to re-optimization (with the MMFF94 force field). Most stable structure for each compound was generated and used for calculating various physicochemical parameters used for the statistical analysis. The resulted geometries were used for docking study.

2.2.2. Calculation of fragment-based descriptors

For the generated descriptors, a pool of about 856 2D-3Ddescriptors was calculated using the PaDEL-Descriptor v2.20 software package. These descriptors include Acidic group count, ALOGP, APol, Aromatic atoms count, BCUT, Chi cluster, constitutional, Eccentric connectivity index, electropological state, XLogP, Zagreb index, Moment of inertia, Zagreb index, Topological charge, Charged partial surface area, Wiener numbers, Petitjean shape index, RDF, WHIM etc. All descriptors with constant values among the dataset were deleted, resulting in 316 different descriptors (independent variables) which were used in the QSAR analysis.

2.2.3. Selection of the training and test sets

In order to compare the biological activities of the set of compounds which have a wide range of chemical structures (i.e., descriptors), the dataset was divided into representative training and test sets using a dissimilarity-based compound selection method called Kennard-Stone algorithms. The program is intended to split a source dataset to training and test sets for further modeling. There are many cases when the splitting to training and test set is complicated because of poor endpoint variables range and etc. In this program authors implemented Kennard-Stone algorithm which takes into account all available information (descriptors) to make a splitting, to get evenly distributed set of data in both sets. The program is very quick, easy to use, with well documented manual that includes background information and steps to run the software. On my opinion the program is very useful and can be applied for many kinds of datasets, which need to be splitted to develop and validate a predictive model.

Table 1 *Structures of dataset used for GA-MLR QSAR analysis with corresponding observed and predicted class of tyrosinase inhibitors.*

Compound ID	Structures of dataset	Observed pIC50	Predicted pIC50	Residual
ID01		-0.816	-1.332	0.515
ID02		-1.425	-1.008	-0.417
ID03		-1.090	-1.200	0.110
ID04		-1.230	-1.308	0.078
ID05		-1.071	-0.776	-0.295
ID06		-0.684	-0.950	0.266
ID07		-1.295	-1.540	0.245
ID09		-0.681	-0.839	0.158
ID10		-0.831	-1.128	0.297
ID11		-0.320	-0.013	-0.307
ID15		-0.417	-0.352	-0.065
ID16		-0.616	-0.713	0.097
ID17		-1.164	-1.315	0.151

ID18	-0.957	-0.577	-0.380
ID19	-0.568	-0.746	0.178
ID20	-1.108	-0.918	-0.190
ID21	-1.186	-1.206	0.020
ID22	-0.819	-0.558	-0.261
ID23	-1.854	-1.913	0.059
ID24	-0.603	0.503	-1.106
ID25	-0.314	0.631	0.945
ID26	-1.127	-0.939	-0.188
ID27	-0.504	-0.773	0.269
ID28	-1.103	-0.894	-0.209
ID29	2.233	2.136	0.097
ID30	1.909	1.807	0.102
ID31	2.000	1.822	0.178

ID32	1.580	2.163	-0.583
ID33	2.097	1.866	0.231
ID34	2.213	2.298	-0.085
ID35	1.613	1.979	-0.366
ID36	2.205	1.948	0.257
ID37	1.940	1.864	0.075
ID38	1.699	1.967	-0.268
ID39	1.000	1.652	-0.652
ID42	1.699	1.598	0.101
ID40	1.177	1.6502	-0.4732

2.2.4. Optimized variable selection

Owed to the fact that it is tedious and unreasonable to investigate all possible combinations of the descriptor pool, genetic function approximation and multiple linear regression, which simplify the process and reduce the time required to execute algorithms, were implemented [17].

2.2.5. Genetic Function Approximation

Genetic Function Approximation (GFA) [18] is used to determine the best initialization of clusters as well as optimization of initial parameters. Genetic Function Approximation attempt to incorporate the ideas of natural evolution [19]. In general, they start with an initial population, and then a new population is created based on the notion of survival of the fittest. Typically, fitness is the measure for how good this population is and can be calculated depending on the nature of the application, where a distance measure is the most common [20]. Then a process called crossover is done over the new population where substrings from selected pairs are swapped [21].

Multiple Linear Regression is a method used for modeling the linear relationship between dependent variable Y (pIC50) and independent variable X (descriptors). MLR is based on the least squares method: the model is fitted such that the sum-of-squares of differences of observed and a predicted value is

Table 2 Internal Validation Parameters

Parameters	Values
SEE	0.2086
R^2	0.9823
R^2 adjusted	0.9777
F	210.92 (DF :5, 19)
Q	4.7512
FIT	21.0889

Table 3 Leave-One-Out (LOO) Result (Without Scaling)

Parameters	Values
Q^2	0.9705
PRESS	1.3774
SDEP	0.2347
rm^2(Loo)	0.9651
rm^2'(Loo)	0.9671
average rm^2(LOO)	0.9661
delta rm^2(LOO)	0.0020

minimized. MLR estimates values of regression coefficients (R^2) by applying least squares curve fitting method. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points. In regression analysis, conditional mean of the dependent variable (pIC50) Y depends on (descriptors) *X*. MLR analysis extends this idea to include more than one independent variable.

Regression equation takes the form:

$$Y = B1 * X1 + B2 * X2 + B3 * X3 + \ldots\ldots\ldots + c \quad (1)$$

where *Y* is dependent variable, '*B*'s are regression coefficients for corresponding '*X*'s (independent variable), '*c*' is a regression constant or intercept [22].

3. Results and Discussions

3.1. Results

All molecules in each data set were successfully optimized by Spartan 14 V1.1.2 software. The following properties were obtained from the optimized structures: Molecular properties, QSAR descriptors, thermodynamic properties as well as acidity and basicity properties. The successful optimization of the molecules implies that all the molecules used have geometries close to their real or test tube geometries. Thus, properties computed from these optimized molecules are reliable.

3.1.1. Descriptor Calculation

The descriptor of each molecular structure was successfully computed with the aid of PaDEL version 2.20 descriptor tool kits. Approximately 856 descriptors ranging from 1D, 2D, and 3D were obtained from these soft ware's.

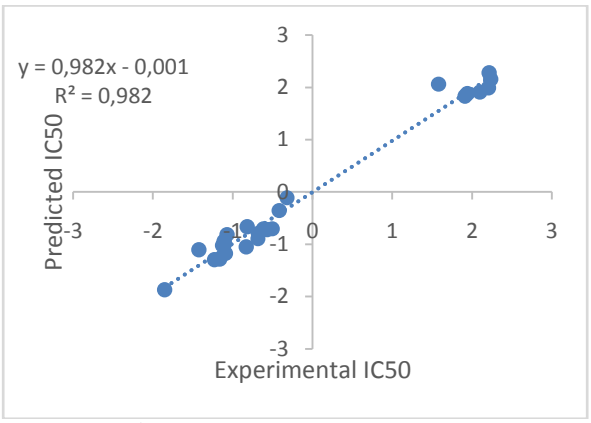

Fig. 1. Scatter plot of the experimental activities versus predicted activities for QSARmodel, LOO cross-validated predictions on full training set.

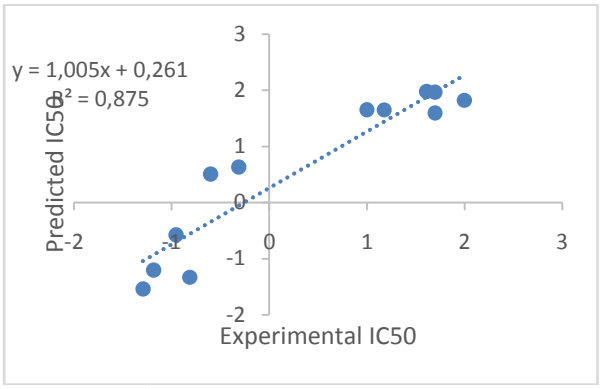

Fig. 2. Scatter plot of the experimental activities versus predicted activities for test-set predictions

3.1.2. GA-MLR Derived models for pIC50 Anti-tyrosinase Compounds

Models 1 give the best Genetic Function Approximation-Multiple Linear Regression (GA-MLR) derived QSAR models for pIC50 of anti-tyrosinase molecules. Based on the model with the best statistical parameters identified using the parameters in Table 6 as standard, Model 1 were chosen as the best models for predicting the pIC50 of anti-tyrosinase molecules. The internal and external validation parameters of the models conform to the minimum standard for a robust QSAR model shown in Table 1-8, confirming the stability and robustness of the models.

3.1.3. Genetic algorithm-multi-parameter linear regression

$$\text{pMIC50} = 1.40103(+/-1.69175) - 0.0257(+/-0.0011)\ \text{ATS0s} - 3.73751(+/-0.98995)\ \text{AATS6p} + 0.19682(+/-0.01267)\ \text{ATSC1i} + 1.17379(+/-0.10375)\ \text{SpMAD\_Dzv} - 0(+/-0)\ \text{VR1\_Dze}.$$

Model 1

Table 4 External Validation Parameters (Without Scaling)

Parameters	Values
r^2	0.8756
r0^2	0.8574
reverse r0^2	0.8406
rm^2(test)	0.7574
reverse rm^2(test)	0.7118
average rm^2(test)	0.7346
delta rm^2(test)	0.0456
RMSEP	0.5389
Rpred^2	0.8323
Q^2_{f1}	0.8323
Q^2_{f2}	0.8121

Table 5 GA-MLR Overall Parameters

Parameters	Values
rm^2(overall)	0.8797
reverse rm^2(overall)	0.8756
average rm^2(overall)	0.8777
delta rm^2(overall)	0.0041

Table 6 Golbraikh and Tropsha (2002) acceptable model criteria's

Parameters	Values
Q^2	0.9705, Passed (Threshold value Q^2 > 0.5)
r^2	0.8756, Passed (Threshold value r^2 > 0.6)
|r0^2-r'0^2|	0.017, Passed (Threshold value |r0^2-r'0^2|<0.3)
k	0.8192, Passed Threshold value: [0.85<k<1.15]
[(r^2-r0^2)/r^2]	0.02, Passed Threshold value: (r^2-r0^2)/r^2) < 0.1
k'	1.0583, Passed Threshold value: 0.85<k'<1.15
[(r^2-r'0^2)/r^2]	0.04, Passed Threshold value: (r^2-r'0^2)/r^2) < 0.1

3.1.4. ***Plot of Experimental Versus Predicted pIC50 of model 1***

The agreement between the experimental pIC_{50} values of molecules used in the training set and the predicted values by the optimization models 1 presented in Figs. 1 and 2, respectively. The high Linearity of these plots indicates high predictive power of the models.

3.1.5. ***Residual plot of model 1***

The measure of the dispersion of residual pIC_{50} values from the predicted pIC_{50} values are presented in Fig. 3. The propagation of the errors on both sides of zero is an indication of the robustness of the QSAR models.

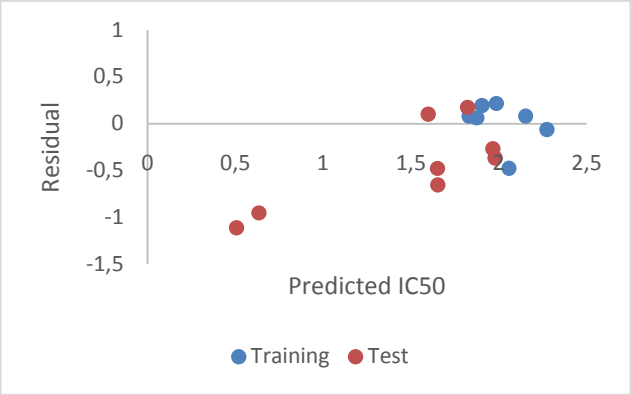

Fig. 3. The residual versus the experimental pIC50 by measured GA-MLR.

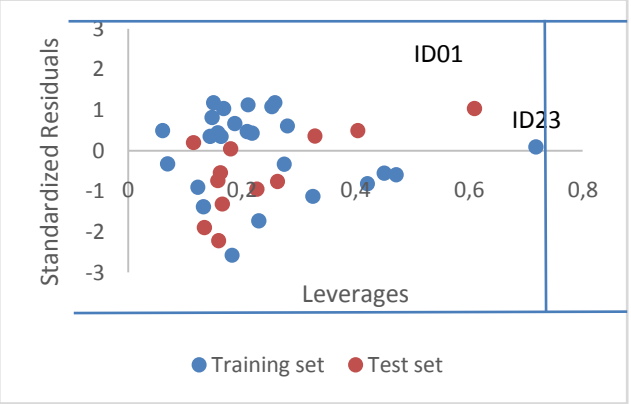

Fig. 4: William's plot of generated GA-MLR model.

3.1.6. ***Comparison of observed and predicted pIC50 of model 1***

The comparison of the predicted pIC_{50} of the model with their experimental values are presented in Tables 1. The low residual values shown in the tables confirms the high predictive power of the models.

3.2. Discussion

After analyzing, we split the data set into the training set and query set, the next step was to select the main factors which were the most important for the anti-tyrosinase inhibition. As we do not know yet which descriptors or which particular combinations are related to the studied response and can be used in the predictive models, we applied genetic algorithms as the variable selection procedure to select only the best combinations (most relevant) for obtaining the models with the highest predictive power by using the training set. Five most significant descriptors according to the GA-MLR algorithm are Broto-Moreau autocorrelation - lag 0 / weighted

by I-state (ATSOs), Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities, (AATS6p), Centered Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential (ATSC1i), Spectral mean absolute deviation from Barysz matrix / weighted by van der Waals volumes (SpMAD_Dzv), and Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities (VR1_Dze). The predicted values for pIC50 for the compounds in the training and test sets using model 1 were plotted against the experimental pIC_{50} values in Figs. 1 and 2. A plot of the residual for the predicted values of pIC_{50} for both training and test sets against the experimental pIC_{50} values are shown in Fig. 3. Clearly, the model did not show any proportional and systematic erroras suggested by Jalali-Heravi and Kyani [23], because the propagation of the residuals on both sides of zero is random. The real usefulness of QSAR models is not just their ability to reproduce known data verified by their fitting power (R^2), but mainly it is their predictive application potential.The *F* -value has found to be statistically significant at 95% level, since the calculated *F* valueis higher as compared to tabulated value. The positive value of quality factor (Q) for this QSAR's model suggests its high predictive power and lack of over fitting [24,25].

A statistically significant 2D-QSAR model was obtained using the properly selected training set of 25 ligands. Results of the statistical analysis are presented in Tables 1 to 8. In the QSAR model, initial GA analysis of the aligned training set was done using material studio version 7.0. This yielded a highly significant Q^2 value of 0.9705 Table 3 (with SDEP= 0.2347 Table 2), which indicates that it is a model with high statistical significance; a Q^2 value of 0.6 is considered statistically significant in QSARstudies [26]. The conventional R^2 value of 0.9823 and low standard error of estimate (SEE) value of 0.2086 Table 1, indicate the accuracy of the predictions of the model. High values of Q2 from the leave-one-out (LOO) analysis (Table 3) can be regarded as a necessary, but not a sufficient, condition for a model to possess significant predictive power [27]. In addition to LOO, the internal predictive ability of the model was further assessed by a Y-randomization performed with 25 analogues for 10 times. The average of 10 readings was given as average Q^2 as showed in Table 7; Y-randomization test (Table 7) ensures the robustness of a QSAR model [28] and to assess the multiple linear regression models obtained by descriptor selection [29]. In y-randomization test, the dependent variable or biological activity is randomly shuffled and a new QSAR model is developed keeping molecular descriptors intact. The new models are expected to have low R^2 and Q^2 values, which determine the statistical significance of the original model. Moreover, if the model development includes *F*-stepping, then it is necessary to shuffle both dependent and independent variables to indicate that the original model is not because of chance correlation. The low R^2 and Q^2_{LOO} values of the random models shown in Table 7 and the value of R^2_p= 0.8988 ($R^2_p \geq 0.5$) indicates that there is no chance of correlation or structural dependency in the proposed model. Consequently model 1 can be considered as a perfect model with both high statistical significant and excellent predictive ability.

FIT Kubinyi function define the statistical quality of activity prediction, the number of variables that enter in a QSAR model are compared by using FIT Kubinyi function as showed in the equation below, a criteria closely related to F value was proven to be useful.

$$FIT = R^2 (n - k - 1) / (n + k^2) (1 - R^2)$$

where *n* is the number of compounds in training set and *k* is the number of variables in the QSAR equation. The main feature of the *F* value is its sensitivity to changes in *k*, if *k* is small sensitivity is high and vice versa if *k* is large. The FIT criterion has a low sensitivity towards changes in *k* values, as long as they are small numbers, and a substantial increase in sensitivity for large *k* values [30,31]. The best model will be the one that possess a high value of this function. According to the statistical values of the models reported in Table 3, with five variables since this showed high FIT. The observed, calculated and predicted values of the statistically significant five parameter QSAR model are presented in model 1.

To satisfy with the robustness of the QSAR model developed using the training set, we have applied the QSARmodel to an external data set of tetraketone and benzyl benzoate derivatives constituting the test set. As the experimental values of IC_{50} for these inhibitors are already available, this set of molecules provides an excellent data set for testing the prediction power of the QSAR model for new ligands. Fig. 2 represents the predicted pIC_{50} values of the test set based on equation (1). The overall root mean square error of prediction (RMSEP) between the experimental and predicted pIC_{50} values

Table 7 *R*, *R^2*, *Q^2* and *Rp^2* values after several *Y*-Randomization test

Model	R	R^2	Q^2
Original	0.9911	0.9823	0.9705
Random 1	0.4627	0.2141	-0.2846
Random 2	0.4213	0.1775	-1.0689
Random 3	0.3681	0.1355	-0.3185
Random 4	0.5475	0.2997	-0.2090
Random 5	0.3707	0.1374	-0.6789
Random 6	0.3160	0.0998	-0.4111
Random 7	0.4059	0.1647	-0.7249
Random 8	0.4484	0.2011	-0.4412
Random 9	0.2769	0.0766	-0.6252
Random 10	0.3824	0.1462	-0.5918
Random Models Parameters			
Average r :		0.3999	
Average r^2 :		0.1652	
Average Q^2 :		-0.5354	
Rp^2 :		0.8988	

was 0.5389 as showed in Table 4, which reveals good predictability. The estimated correlation coefficients between experimental and predicted pIC_{50} values with intercept (*r0^2*) and without intercept (*r^2*) were 0.8406 and 0.8574, respectively. The value of [(*r^2 –r0^2*)/*r^2*] = 0.002 (Table 6), which is less than 0.1 stipulated value [32]and thus validates the usefulness of the QSAR model for predicting the biological activity of the external data set. Also, the values of *k* and *k′* were 0.8192 and 1.0583, which are well within the specified ranges of 0.85 and 1.15 [28].The values of R^2pred= 0.8323 and *rm^2*(test) = 0.7574 were found to be in the acceptable range (Table 4) [33], there by indicating the good external predictability of the QSAR model.

Selecting the best model, values of *rm^2*(overall) for the model was determined. As shown in Table 5, thisparameter penalized a model for large differences in experimental and predicted activity values. The parameter *rm^2* (overall) determines whether the predicted activities are really close to the observed values or not since high values of Q^2 and R^2 preddoes not necessarily mean that the predicted values are very close to the experimental ones. A model is considered satisfactory when *rm^2* (overall) is greater 0.5 [34]. Besides *rm^2* (overall), we have calculated *rm^2* (test) and *rm^2* (LOO) values Tables 3 and 4. These two parameters signify the differences between the experimental and predicted activities of the test and training set compounds. For an ideal predictive model, the difference between R^2pred and *rm^2* (test) in Table 4 and difference between Q^2 and *rm^2* (LOO) Table 3 should be low. Large difference between the values will ultimately lead to poor values of *rm^2* (overall) parameter. For this data set, the difference between Q2and *rm^2*(LOO) is quite less (0.0054) and that between R2predand rm^2(test) is also very less (0.0749). Thus indicates that the model obtained for this data set using those descriptors are quite robust and predictive. The *rm^2*(LOO) parameter in Table 3 for a given model indicates the extent of deviation of the LOO predictedactivity values from the experimental ones for the training set compound while parameter *rm^2*(test) (Table 4) determines the extent of deviation of the predicted activity from the experimental activity values of test set compounds where the predicted activity is calculated on the basis of the model developed using the corresponding training set. Model 1 show acceptable values of *rm^2*(LOO) and *rm^2*(test) since they are greater than 0.5 [28].

The multi-collinearity between the above five descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as:$1/1 - R^2$ [35].

Where *r* is the correlation coefficient of the multiple regression between the variables in the model. If VIF equals 1, no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [26]. The corresponding VIF values of the seven descriptors are shown in Table 5. Based on this table, most of the variables had VIF values of less than 5, indicating that the obtained model has statistical significance. To examine the relative importance, as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed using the following equation.

$$MF_j = \frac{B_j \sum_{i=1}^{j=n} d_{ij}}{\sum_j^m B_j \sum_i^n d_{ij}} \quad (1)$$

3.2.1. *Applicability domain of the model*

A quantitative structure activity relationship (QSAR) model is exploited to monitor new compounds when its domain of application has been defined [28]. The prediction may be assumed reliable for only those compounds which fall into this domain [36]. Standardizedresiduals of the activity were computed and were plotted versus leverage values (*h*). The value of leverage was calculated for every compound. Values are always between 0 and 1. A value of 0 is indicative of perfect prediction and usually is not accessible, and a value of 1 indicates very poor prediction. The lower the value, the higher confidence in the prediction. Warning leverage (*h**) is another standard for explanation of the results and is, generally, fixed at 3 (*k +1)/ n*, where *k* is the number of model parameters and n is the number of training and test sets [36]. Calculated leverage for training and test sets is useful for determining the compounds which affect the model and, in terms of validation set, useful for assigning the applicability domain of the model. The William's plot for the developed models in GA-MLR are shown in Fig. 3. Response outliers are compounds that have standard residual points higher than ± 3.0 standard deviation units and a leverage value higher than the warning leverage, which is 0.72 for GA-MLR. As can be seen in Fig. 3, all studied molecules in training and test sets lie with high degree of confidence in application domain of the developed models.

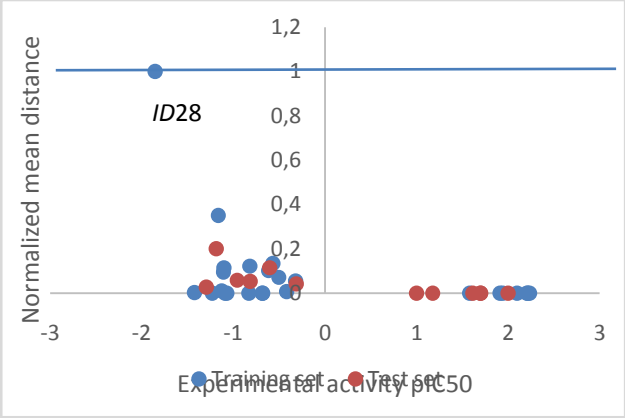

Fig. 5. Euclidean based applicability domain generated GA-MLR model

Table 8: The linear model based on the five parameters selected by the GA-MLR method.

Descriptors name	Symbol	VIF	MF
Broto-Moreau autocorrelation - lag 0 / weighted by I-state	ATS0s	1.5719	5.1762
Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities	AATS6p	1.3968	3.2661
Centered Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential	ATSC1i	1.3411	-0.7138
Spectral mean absolute deviation from Barysz matrix / weighted by van der Waals volumes	SpMAD_Dzv	1.7905	-6.8643
Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities	VR1_Dze	1.9766	0.1357

3.2.2. Interpretation of descriptors

The 2D-QSAR developed indicated that Broto-Moreau autocorrelation - lag 0 / weighted by I-state (ATSOs), Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities (AATS6p) and N Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities (VR1_Dze) has positive values in the mean effect (Table 8) indicate that the indicated descriptor contributes positively to the value of pIC_{50}, whereas negative values of Centered Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential (ATSC1i), and Spectral mean absolute deviation from Barysz matrix / weighted by van der Waals volumes (SpMAD_Dzv) indicate that the greater the value of the descriptor the lower the value of pIC_{50}. In other words, increasing the ATSC1i and SpMAD_Dzv (Table 8) will decrease pIC50 and increasing the ATS0s, AATS6p and VR1_Dze increases extent of pIC_{50} of the tetraketone and benzyl benzoate derivatives. The mean effect reveals the significance of an individual descriptor presented in the regression model.

4. Conclusion

The generated 2D-QSAR equations, described in this publication, indicate the relationship between biological activity and the corresponding descriptors. Table 8 presents abbreviations, full names and description of all the descriptors used in the final 2D-QSAR models. Some of them have quite clear physical meaning, but unfortunately, most of them is often “a combination of” a few physical and chemical properties. Not often happens that the biological activity of the drug is dependent on one or a few obvious and clear properties. Many factors have an influence on the biological activity of the compound and getting to know them isn’t easy. Time consuming and costly researches of professionals in the drug design confirm this facts. Obtained 2D-QSAR models allow to predict the activity of a new compound on the basis of its structure without the need of its synthesis. Estimation of the predictable (predicted) biological activity for the next analog can be made by optimizing the geometry of the compound in a suitable computer program and using appropriate computational quantum chemistry methods and then compute the so-called molecular descriptors. At that time, the obtained values of descriptors are substituted for the found earlier, reliable 2D-QSAR models and used to calculate a predicted value for the biological activity of a new tetraketone and benzyl benzoate derivatives of the analyzed group of compounds. The predictive ability of the model and the internal and external validation procedures illustrated the accuracy on one hand and offered a useful alternative to the time consuming experiments, on the other. In order to propose structural modifications that can be taken into account in the further synthesis of next analogues, we plan to carry out a molecular docking. It involves the generation, for a series of compounds, so called molecular field, which allows to visual identification of areas with positive or negative impact on the biological activity. This work emphasizes the use of various tests in QSAR analysis such as applicability domain of the model, predictive ability of validation set and FIT Kubinyi function as important parameters to obtain a reliable and robust QSAR model and thus help in designing more potent anti-tyrosinase inhibitors.

Recommendation

1 These drugs like molecules may be synthesized and formulated appropriately.
2 Their pharmacological and toxicological activities could be performed on animal models before clinical trials.

References

[1] Abechi SE and Edache EI. Application of Genetic Algorithm-Multiple Linear Regression (GA-MLR) For Prediction of Anti-Fungal Activity. Inter J of Pharma Sci and Res 2016; 7: 204- 20.
[2] Seo SY, Sharma VK, and Sharma N, Mushroom tyrosinase: Recent prospects, J. Agric. Food Chem 2003:51; 2837–53.
[3] Khan MTH. Molecular design of tyrosinase inhibitors: A critical review of promising novel inhibitors from synthetic origins. Pure Appl. Chem , 2007:12; 2277–2295

[4] Khan KM, Maharvi GM, Khan MTH, Shaikh AJ, Perveen S, Begum S, Choudhary MI. Tetraketones: A new class of tyrosinase inhibitors Bioorganic and Med Chem 2006:14; 344–51
[5] Shiino M, Watanabe Y, Umezawa K. Synthesis of N-substituted N-nitrosohydroxylamines as inhibitors of mushroom tyrosinase. Bioorg. Med. Chem. 2001:9;1233-1240.
[6] Kanchanapally RC, Macha R, Vunguturi S. and Tigulla P. A Theoritical Study of Benzyl Benzoates with Agaricus Bisporus Tyrosinase Inhibitory Properties. Inter J of Life Sci and Pharma Res 2011: 1; 28 -40.
[7] Lee TH, Seo JO, Baek S,and Kim SJ. Inhibitory Effects of Resveratrol on Melanin Synthesis in Ultraviolet B-Induced Pigmentation in Guinea Pig Skin. Biomol Ther (Seoul). 2014:22(1); 35–40.
[8] Kim YJ, Uyama H. Tyrosinase inhibitors from natural and synthetic sources: structure, inhibition mechanism and perspective for the future. Cell Mol Life Sci. 2005:62(15); 1707-23.
[9] Cramer, RD. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. J. Med. Chem. 2003:46; 374–388.
[10] Giersiefen H, Hilgenfeld R, Hillisch A, Modern Methods of Drug Discovery: An Introduction. In Modern Methods of Drug DiscoVery; Hilgenfeldl, A. H. R., Ed.; Birkha¨user Verlag: Basel, 2003; 1-18.
[11] Ruchi RM, Ross AM, and Michael JS. Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization. J. Chem. Inf. Model2009:49;1810–1820.
[12] Sprous DG, Palmer RK, Swanson JT, Lawless M. QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. Curr Top Med Chem. 2010; 10(6):619–637.
[13] Wavefunction, Inc. Spartan'14, version 1.1.2, Irvine, California, USA; 2013.
[14] Dewar MJS, Zoebisch EG, Healy EF & Stewart JJP. The development and use of quantum mechanical molecular models. 76. AMI: a new general purpose quantum mechanical molecular model. J. Amer. Chem. Soc. 1985:107;3902-9.
[15] Stewart JJP. Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. J Mol Model 2004: 10;155–164.
[16] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J of Computational Chem. 2011:32(7); 1466-1474.
[17] KirkpatrickS, Gelatt CD, Jr.Vecchi MP.Optimization by Simulated Annealing.Science, New Series 1983:220;671-80.
[18] Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J of Chem Inf and Comput Sci. 1994: 34; 854-66.
[19] Ching-Wen H, Lin KP, Wu MC, Hung KC, Liu G, and Jen C. Intuitionistic fuzzy c-means clustering algorithm with neighborhood attraction in segmenting medical image. Soft Computing. 2014: 1-12.
[20] Arici T, Celebi S, Aydin AS, Temiz TT. Robust gesture recognition using feature pre-processing and weighted dynamic time warping. Multimedia Tools and Applications 2013;1-18.
[21] Sumathi P and Kathiresan V. A Hybrid Model for Medical Data Using Machine Learning Approaches. Inter J of Mod Trends in Engr and Res.2016:2349-9745.
[22] Thakur M, Thakur A, Ojha L. Surface Area Grid in Modeling of Anti HIV Activity of TIBO Derivatives. Inter Jof Res and Development in Pharm and Life Sci. 2014:3(3); 983-992.
[23] Jalali-HeraviM. andKyani A. Use of Computer-Assisted Methods for the Modeling of the Retention Time of a Variety of Volatile Organic Compounds: A PCA-MLR-ANN Approach. *J Chem Inf Comput Sci. 2004:*44 (4); 1328–35.
[24] Pogliani L. "Structure property relationships of amino acids and some dipeptides," Amino Acids 1994:6; 141–153.
[25] Pogliani L. "Modeling with special descriptors derived from a medium-sized set of connectivity indices," J of PhysChem. 1996:100; 18065–18077.
[26] Böhm M, Strzebeche J, Klebe G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. J Med Chem.1999:42; 458–477.
[27] Hattotuwagama CK, Doytchinova IA, Flower DR. In silico prediction of peptide binding affinity to class i mouse major histocompatibility complexes: A comparative molecular similarity index analysis (CoMSIA) Study. J Chem Inf Model.2005, 45, 1415–1423.
[28] Tropsha A, Gramatica P, Gombar VK. The importance of being earnest:validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci. 2003:22;69–77.
[29] Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models J Med Chem. 2005;48 (3):661-663.
[30] Kubinyi H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. Quant Struct Act Relat. 1994;13:393-401.
[31] Kubinyi H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. Quant Struct Act Relat. 1994;13:285-294.
[32] Edache EI, Hambali UH, Arthur DE, Oluwaseye A and Chinweuba OC. In-silico Discovery and Simulated Selection of Multi-target Anti-HIV-1 Inhibitors. Inter Res J of Pure and Applied Chem. 2016:11(1); 1-15.
[33] Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci. 2008:27; 302-313.
[34] Golbraikh A, Tropsha A. Beware of q2! J Mol Graphic and Model. 2002:20; 269-276.
[35] Edache EI, Uzairu Aand Abechi SE. Quantitative structure and activity relationship modeling study of anti-HIV-1RT inhibitors: Genetic function approximation and density function theoryMethods. J of Comput Mthds in Mol Design.2015:5 (4);61-76.
[36] Mansourian M, Fassihi A, Saghaie L, Madadkar-Sobhani A, Mahnam K, Abbasi M. QSAR and docking analysis of A2B adenosine receptor antagonists based on non-xanthine scaffold. Med Chem Res. 2015;24:394–407.