Quantitative Structure and activity Relationship of 3a, 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives as anti HIV-1 Agents

Ahanonu Saviour UGOCHUKWU a,1, Gideon Adamu SHALLANGWA a, Adamu UZAIRU a

aDepartment of Chemistry, Ahmadu Bello University, Zaria- Nigeria

Abstract: A novel series of 3a, 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives have been reported as better anti-HIV 1 agents. In this study QSAR was carried on a 3a, 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives as anti HIV – 1 agents. Two different variable selection approaches namely: Genetic function approximation and multi linear regression models were used to predict the HIV-1 inhibition activity. The following were obtained after the model was internally validated: squared correlation coefficient (R^2) of 0.8823, adjusted squared correlation coefficient (R^2_adj) of 0.8528 and leave one out (LOO) cross validation coefficient (Q^2_cv) of 0.7566. The external validation was carried out to confirm the predictive power of the model and R^2pred of 0.6901 was obtained. The validated model result above showed that the five descriptors which are GATS6c, VR3_Dze, minHCsats, RDF30m and Eze contributed positively to the power of the model and R^2_cv of 0.7566. The external validation was carried out to confirm the predictive power of the model and R^2pred of 0.6901 was obtained. The validated model result above showed that the five descriptors which are GATS6c, VR3_Dze, minHCsats, RDF30m and Eze contributed positively to the activity. The result obtained will be very helpful for designing and synthesizing other derivatives with improved anti-HIV activities.

Keywords: HIV, AIDS, QSAR, 3a 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives, model validation.

1. Introduction

Human immunodeficiency virus type 1(HIV-1) is the main causative agent of acquired immunodeficiency syndrome (AIDS) which remains a serious public health problem throughout the world [1]. HIV-1 integrase (IN) is a virally encoded enzyme essential for virus replication, which mediates insertion of the double-stranded DNA provirus into the host genome[2]. Integration is the final step before irreversible and productive HIV-1 infection of the target cell [3]. During the past two decades an increasing number of quantitative structure-activity/property relationship (QSAR/QSPR) models have been studied using theoretical molecular descriptors for predicting biomedical, activity, toxicology and technological properties of chemicals.

QSAR was performed on 3a, 6a – Dihydro-1H-pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives dataset. The overall goals of QSAR retain their original essence and remain focussed on the productive ability of the approach and its receptiveness to mechanistic interpretation. QSAR includes all statistical methods by which biological activities (most often expressed by logarithms of equipotent molar activities) are related with structural elements, physiochemical properties or fields (3D QSAR) [4]. Following our interest in this field, our aim is to describe the structure-activity relationships study on 3a, 6a – Dihydro-1H-pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives and develop a QSAR model on these compounds with respect to their 50% effective concentration(EC_{50}).

2. Materials and Methods

The experimental effective concentrations (EC_{50}) in micromole of 3a, 6a – Dihydro-1H-pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives against HIV-1 integrase inhibitors are extracted from a recent publication[5]. For modelling purposes these values are converted into logarithm units (-log_{10}EC_{50}). Table 1 shows the experimental activities in Log EC_{50} of 3a, 6a – Dihydro-1H-pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives. The

---

1 Corresponding Authors

e-mail: favour_saviour@yahoo.com
A dataset of 35 compounds were divided into 26 training sets to build the model and 9 test sets to validate the model.

Structure of 3a, 6a – Dihydro-1H- pyrrolo[ 3,4-c] pyrazole-4,6-dione Derivatives

![Structure of 3a, 6a – Dihydro-1H- pyrrolo[ 3,4-c] pyrazole-4,6-dione Derivatives](image)

**Figure 1.** Compounds 1-20

**Table 1.** 3a, 6a – Dihydro-1H- pyrrolo[ 3,4-c] pyrazole-4,6-dione Derivatives and their respective activities.

| Comp No | R             | LogEC50 |
|---------|---------------|---------|
| 1*      | 2,3-OHPh      | 4.5786  |
| 2       | 2-OMe, 3-OHPh | 5.2700  |
| 3       | Ph            | 5.0357  |
| 4*      | 2-OHPh        | 4.9952  |
| 5       | 2,4-OHPh      | 4.4521  |
| 6*      | 2-OH, 3-FPh   | 5.2411  |
| 7*      | 2-OH, 5-FPh   | 5.3757  |
| 8       | 2-OH, 3-ClPh  | 4.6899  |
| 9       | 2-OH, 3-FPh   | 5.1221  |
| 10      | 2-OH, 3-NO2Ph | 5.3098  |
| 11      | 2-OH, 3-OMePh | 5.4001  |
| 12      | 2-OH, 4-OMePh | 5.0696  |
| 13      | 2-OH, 5-OMePh | 5.3251  |
| 14      | 2-OH, 3-OEtPh | 3.9360  |
| 15      | 2-OH, 3-OMe, 5-NO2Ph | 4.3774 |
| 16      | 2,3-OMePh     | 4.9535  |
| 17*     | Benzo[1,3]dioxol-4-yl | 5.1481 |
| 18      | 2-OH-naphthalene-1-yl | 5.3468 |
| 19      | Thiazol-2-yl  | 4.4492  |
| 20      | Pyridine-2-yl | 4.5824  |
Test set compounds are represented with.

2.1. Optimization

The structures of all the compounds were drawn using ChemDraw Ultra module. The drawn structures were imported to Spartan 14 where the 3D structures of the 35 compounds were created. Their energies were minimized by molecular mechanics force fields (MMFF) to remove the strain energy before subjecting it to quantum chemical estimations. DFT (Density Functional Theory) with B3LYP (6-311G*) basis set was employed for complete optimization. The Spartan files of all the optimized molecules were then saved in the SD file format which is the recommended input format in PaDEL Descriptor software V2.20 [6]. The optimization was carried out using Spartan 14.

2.2. Molecular Descriptor Calculations

Descriptors are mathematical values used to describe the properties of molecules. The 35 compounds descriptors calculation was calculated using PaDEL-Descriptors software V2.20. A total of 1629 molecular descriptors were calculated.

2.3. Normalization of descriptors

The descriptors’ value was normalized using Equation 1 in order to give each variable the same opportunity at the onset to influence the model [7].

\[
X = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \quad (1)
\]

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 respectively.

2.4. Data Pretreatment

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 [6].

2.5. Data Division

Data Division software obtained from Drug Theoretical and Cheminformatics Laboratory (DTC Lab) by employing Kennard and Stone’s algorithm was used in order to obtain validated QSAR models from the dataset. The dataset was divided into 26 training and 9 test sets in the percentage of 75% and 25% respectively which table 1 clearly shows.

2.6. Model Validation

Validation of the model was performed using Material studio software version 8 by utilizing Genetic Function Approximation (GFA) method. The importance of model validation could now be regarded as a collective wisdom within the community of molecular modellers [8].

LOF (Friedman’s lack of fit) was one of the methods used to validate the model. The formula is given in equation 2 below.

---

Table 2. 3a, 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives and their respective activities.

| Comp No | R₁ | R₂   | LogEC₅₀ |
|---------|----|------|---------|
| 21      | 4F | 3F   | 4.9179  |
| 22      | 4F | 2F   | 3.9475  |
| 23      | 4F | 4Cl  | 4.5940  |
| 24      | 4F | 2.4-F | 4.1810 |
| 25      | 4F | 4CF₃ | 4.4935  |
| 26      | 4F | 4-SO₂Me | 4.1848 |
| 27      | 4F | 4-SO₂NH₂ | 4.5167 |
| 28*     | 4F | H    | 4.5432  |
| 29      | 4F | 4-Me | 5.3872  |
| 30      | 4F | 4-OMe | 4.6753 |
| 31      | 2F | 4F   | 5.5986  |
| 32      | 3F | 4F   | 4.8854  |
| 33*     | 4Cl | 4F  | 4.9817  |
| 34*     | H  | 4F   | 4.7242  |
| 35*     | 4-OMe | 4F | 4.0597  |

*Test set compounds are represented with.

---
where SEE is the standard error of estimation, \( c \) is the number of descriptors, \( p \) is the number of independent parameters, \( m \) is the number of samples and \( d = 1 \). The advantage of using LOF rather than SSE is that LOF do not decrease with increase in the number of descriptors. The lower value of LOF in QSAR indicates that the model has a good predictive power.

The second parameter is cross- validation which is based on leave one out (LOO) or leave some out (LSO) cross validation procedure. The outcome from this procedure is the cross-validation parameters. They include PRESS (predicted residual sum of squares), \( SSY \) (sum of the squares of the response values), \( S_p \) (uncertainty of precision), \( Q^2_{cv} \) overall predicted ability and PSE (predictive square Error). Frequently \( Q^2_{cv} \) is used as a criterion of both robustness and predictive ability of the model. High value of \( Q^2_{cv} \) (for instance >0.5) is an indicator of the high predictive power of the QSAR model.

\[
Q^2_{cv} = 1 - \frac{\sum(Y_{cat} - Y_{obs})^2}{\sum(Y_{obs} - \bar{Y})^2}
\]  

(3)

Correlation coefficient between the predicted and observed activities, \( R^2 \) is the third parameter for validating a model but not a complete useful measure of stability of a model. \( R^2 \) varies directly with the increase in number of descriptors.

\[
R^2 = 1 - \frac{\sum(Y_{obs} - Y_{cal})^2}{\sum(Y_{obs} - \bar{Y})^2}
\]  

(4)

\( Y_{obs}, Y_{cal} \) and \( \bar{Y} \) are the observed activity, the calculated activity and the mean observed activity of the samples in the training set, respectively.

Another parameter is adjusted squared correlation coefficient (\( R^2_{adj} \)). The formula for calculating \( R^2_{adj} \) is:

\[
R^2_{adj} = \frac{R^2 - (p/n-1)}{1-(1-R^2)/n-p+1}
\]  

(5)

\( P \) in equation 5 is the number of independent variables in the model.

The coefficient of determination of the test set was calculated with the formula in equation (6) below.

\[
R^2_{predicted} = \frac{\sum(y_{pred_{test}} - y_{exp_{test}})^2}{\sum(y_{exp_{test}} - \bar{y})^2}
\]  

(6)

2.7. Y Randomization

Y randomization is carried out only with training set compounds to guarantee the created QSAR model is strong and not inferred by chance. It was carried out by randomly shuffling the dependent variable while keeping the independent variables unaltered. The dependent variable is the activity while the independent variable is the descriptor. The randomized \( R^2 \) and \( Q^2 \) obtained must have lower values after several trials than the original \( R^2 \) and \( Q^2 \) to confirm that the model developed is robust.

Coefficient of determination for Y-Randomization, \( cR^2_{adj} \) must be greater than 0.5 for passing this test [9].

### Table 3. Summary of GFA Analysis

| Analysis type | Genetic Function Approximation |
|--------------|--------------------------------|
| Response column | BJR: activity |
| Number of rows in model population | 26 |
| Maximum generations | 1000 |
| Initial terms per equations | 2000 |
| Maximum length of equation | 5 |
| Length of constant equation | Yes |
| Number of top models returned | 4 |
| Scoring Function | Friedman LOF |
| Scaled LOF smoothness parameter | 0.50000000 |
| Mutation probability | 0.10000000 |
| Linear spine | No |
| Quadratic spine | No |
| Random number seed | 9999 |
| Minimum prediction fraction for term inclusion | 1.000000e-004 |
| Number of variables requested for plot | 5 |

### 3. Results and Discussion

A QSAR examination was performed to investigate the structure Activity relationship of 35 compounds as potent Anti-HIV 1. In order to assemble a good QSAR model for anti-HIV a decent predictive power Kennard-stone was used to divide the data set into a training set of 26 compounds which was used to develop the model and a test set of 9 compounds which was used to utilize the predictive ability of the built model. Table 4a and 4b below show the experimental, predicted and residual values for 3a, 6a – Dihydro-1H- pyrrolo[ 3,4-c] pyrazole-4,6-dione Derivatives. The low residual
values between the experimental and the predicted activity show that the model is of high predictability.

**Table 4a.** Experimental, Predicted and Residual values of training set of 3a, 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives.

| S/N | Experimental | Predicted values | Residual |
|-----|--------------|------------------|----------|
| 2   | 5.3098       | 5.363592         | -0.05379|
| 3   | 5.4001       | 5.391676         | 0.008424|
| 5   | 5.0696       | 5.212068         | -0.14247|
| 8   | 5.3251       | 5.336812         | -0.01171|
| 9   | 3.936        | 3.918456         | 0.017544|
| 10  | 4.3774       | 4.505073         | -0.12767|
| 11  | 4.39535      | 4.453602         | -0.05825|
| 12  | 5.3468       | 5.160046         | 0.186754|
| 13  | 4.4492       | 4.213953         | 0.235247|
| 14  | 5.27         | 5.057108         | 0.212892|
| 15  | 4.5824       | 4.672004         | -0.0896 |
| 16  | 4.9179       | 5.087736         | -0.16984|
| 18  | 3.9475       | 4.184707         | -0.23721|
| 19  | 4.594        | 4.64802          | -0.05402|
| 20  | 4.181        | 4.128223         | 0.052777|
| 21  | 4.4935       | 4.466479         | 0.027021|
| 22  | 4.1848       | 4.339639         | -0.15484|
| 23  | 4.5167       | 4.182111         | 0.334589|
| 24  | 5.3872       | 5.086451         | 0.300749|
| 25  | 5.0357       | 5.233731         | -0.19803|
| 26  | 4.6753       | 4.852069         | -0.17677|
| 27  | 5.5986       | 5.331183         | 0.267417|
| 29  | 4.8854       | 4.852921         | 0.032479|
| 30  | 4.4521       | 4.597567         | -0.14547|
| 31  | 4.6899       | 4.663171         | 0.026729|
| 32  | 5.1221       | 5.205051         | -0.08295|

**Table 4b.** Experimental, Predicted and Residual values of test set of 3a, 6a – Dihydro-1H- pyrrolo[3,4-c] pyrazole-4,6-dione Derivatives.

| S/N | Activity | Predicted values | Residual |
|-----|----------|------------------|----------|
| 1   | 4.5786   | 5.301255         | -0.72266|
| 4   | 5.1481   | 5.315982         | -0.16788|
| 6   | 4.5432   | 4.83087          | -0.28767|
| 7   | 4.9817   | 4.694192         | 0.287508|
| 17  | 4.7242   | 4.95784          | -0.23364|
| 28  | 4.0597   | 3.703689         | 0.356011|
| 33  | 4.9952   | 4.988736         | 0.006464|
| 34  | 5.2411   | 5.729365         | -0.48827|
| 35  | 5.3757   | 5.577769         | -0.20207|

**Table 5.** Validation parameters from material studio.

|                     | Equation 1 | Equation 2 | Equation 3 | Equation 4 |
|---------------------|------------|------------|------------|------------|
| Friedman LOF        | 0.15497    | 0.156554   | 0.157166   | 0.158315   |
| R-squared           | 0.882272   | 0.881068   | 0.880603   | 0.879731   |
| Adjusted R-squared  | 0.852839   | 0.851335   | 0.850754   | 0.849663   |
| Cross validated R-  | 0.756607   | 0.781141   | 0.725727   | 0.789124   |
| squared             |            |            |            |            |
| Significant Regression | Yes       | Yes        | Yes        | Yes        |
The Genetic Algorithm - Multi linear Regression (GA-MLR) study led to the selection of five descriptors which were used to assemble a linear model for calculating predictive activity on HIV-1. Four QSAR model were built but only the first was used due to statistical significance. The parameter of model 1 which was $R^2$ predicted was calculated. The validation parameters in Table 5 above were in agreement with the threshold value reported in Table 6. It showed that the model was stable and robust.

**Table 6.** Minimum recommended values of validation parameters for a generally acceptable QSAR model

| Name                  | Symbol | Value           |
|-----------------------|--------|-----------------|
| Coefficient of $R^2$  | $R^2$  | $\geq 0.6$      |
| Confidence interval at 95% confidence level | $P(95\%)$ | $< 0.05$ |
| Difference between $R^2$ and $Q^2$ | $R^2 - Q^2$ | $\leq 0.3$ |
| Cross validation $Q^2$ | $Q^2$  | $\geq 0.6$      |
| Minimum number of external test set | $N_{\text{ext.test set}}$ | $\geq 0.5$ |
| Coefficient of $cR^2_p$ for Y Randomization | $cR^2_p$ | $> 0.5$ |

The model number 1 used is:

$$pEC_{50} = 3.101882593 \times \text{GATS6c} - 0.185597104 \times \text{VR3_Dze} + 4.934195547 \times \text{minHCsats} - 0.157014990 \times \text{RDF30m} + 8.505034001 \times \text{E2e} - 0.318780476$$

**Table 7.** Pearson’s correlation for descriptors used in the QSAR optimization model

| Name  | GATS6c | VR3_Dze | minHCsats | RDF30m | E2e |
|-------|--------|---------|-----------|--------|-----|
| GATS6c | -0.062 | 1       |           |        |     |
| VR3_Dze | 0.185 | 0.040  | 1         |        |     |
| minHCsats | 0.220 | 0.030  | 0.934 | 1     |     |
| RDF30m | 0.0312 | -0.155 | -0.786 | -0.736 | 1   |
| E2e   | -0.189 | -0.308 | -0.810 | -0.784 | 0.792 | 1   |

The correlation shown in Table 7 above was an indication that the five descriptors used in the QSAR optimization model do not show high correlation. The $Y$-randomization in table 8 below with $cR^2_p > 0.5$ shows that QSAR model is strong and not inferred by chance. It is also in agreement with the threshold values in Table 6.

**Table 8.** Y-Randomization

| Model | R     | $R^2$ | $Q^2$ |
|-------|-------|-------|-------|
| Original | 0.821036 | 0.674101 | 0.342265 |
| Rand. 1 | 0.383917 | 0.147392 | -2.05655 |
| Rand. 2 | 0.283951 | 0.080628 | -1.49508 |
| Rand. 3 | 0.453379 | 0.205553 | -9.82142 |
| Rand. 4 | 0.455922 | 0.207865 | -0.34115 |
| Rand. 5 | 0.331781 | 0.110078 | -2.97162 |
| Rand. 6 | 0.389811 | 0.151952 | -1.87279 |
| Rand. 7 | 0.419556 | 0.176027 | -5.89422 |
| Rand. 8 | 0.362969 | 0.131746 | -1.02703 |
| Rand. 9 | 0.453342 | 0.205519 | -4.73414 |
| Rand. 10 | 0.502091 | 0.252096 | -6.46537 |

Random Models Parameters

| Average $r$ | 0.403672 |
|-------------|----------|
| Average $R^2$ | 0.166886 |
| Average $Q^2$ | -3.66794 |
| $cR^2_p$ | 0.586998 |
design other 3a, 6a – Dihydro-1H- pyrrolo[ 3,4-c] pyrazole-4,6-dione Derivatives with improved anti-HIV 1 activity.

**Acknowledgement**

We wish to thank everyone who contributed in one way or the other for the success of this work. Their pieces of advice, encouragement and ceaseless prayers are appreciated.

**References**

[1] P. Zhan, C. Pannecouque, E. X. De Clercq, Anti-HIV drug discovery and development: current innovations and future trends. J. Med. Chem. 59 (2016) 2849-2878.

[2] R. Di Santo, Inhibiting the HIV integration process: past, present and future. J. med. chem. 51 (2014) 539-566.

[3] C. M. Farnet, B. Wang, L. Russell, F. D. Bushman, Differential inhibition of HIV-1 preintegration and purified integrase protein by small molecules. Proc. Natl. Acad. Sci. USA 93 (1996) 9742-9747.

[4] V. Ravichandran, R. Harish, J. Abhishek, S. Shalini, P. V. Christapher, A. K. Ram, Validation of QSAR models-Strategies and importance, (2011) 511-519

[5] Guan-Nan Liu, Rong-Hua Luo, Yu Zhou, Xing- Jie Zhang, Jian Li, Liu- Meng Yang, Yong- Tan Zheng and Hong Liu. Synthesis and Anyi-HIV -1 Activity Evaluation for Novel 3a, 6a – Dihydro-1H- pyrrolo[ 3,4-c] pyrazole-4,6-dione Derivatives. (2016).

[6] E.A. Shola, S.A. Uba, A. Uzairu, A novel QSAR model for the evaluation and prediction of (E)- N'- Benzylidenesonicotinohydrazide Derivatives as the potent Anti-mycobacterium Tuberculosis Antibiotics using Genetic Function Approach. Physical Chemistry Research, 6 (2018) 479-492.

[7] P. Singh, Quantitative Structure – Activity Relationship study of subsisted – [1,2,4] oxadiazoles as s1p1 Agonists. J. of current Chemical and pharmaceutical series. (2013).

[8] A. Tropsha, Best practices for QSAR model Development, Validation and Explication. Mol. Inf. 29 (2010) 476-488.

[9] E.A. Shola, E.A. Kalen, A. Mustapha, A.Y. Mahmoud, D. Danzarami, Genetic Function Approximation and Multilinear Regression Approach for Activity modelling of ciprofloxacin Derivatives as potential Anti-prostate cancer Agents: A Theoretical Approach. Kenkyu Journal of pharmacy and Health care. 4 (2018) 6-16.