Asian Journal of Chemistry; Vol. 23, No. 4 (2011), 1857-1860

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in

Anti-HIV Activity Study by Classical QSAR Method for 1-Alkoxymethyl-5-alkyl-6-naphthylmethyl Uracils as HEPT Analogues

UTTAM K. TRIPATHI and I.P. PANDEY*

Department of Chemistry, D.A.V. (Post Graduate) College, Dehradun-248 001, India

*Corresponding author: E-mail: uktripathi@hotmail.com

(*Received*: 12 November 2010; *Accepted*: 27 December 2010) AJC-9433

2D QSAR model have been developed to estimate and predict anti HIV activities against HIV-1 for α and β forms of HEPT analogues (1-alkoxymethyl-5-alkyl-6-naphthylmethyl uracils). First, 15 HEPT analogues of α form were studied, and then 10 HEPT analogues of β form were studied, both studies was having good statistical significance, separately. But, for better understanding of the model, both the studies were harmonized. The conclusion for harmonised study upon a series of 15 HEPT ligands, inhibitors of HIV reverse transcriptase; using the QSAR that imply analysis of correlation and multi linear regression; a significant collection of descriptors was used. The best QSAR model with good correlation coefficient ($r^2 = 0.792$), of high statistical significance (> 99.9 %) well explained the variance in activity.

Key Words: 2D QSAR, HEPT, α and β forms harmonization, Anti HIV.

INTRODUCTION

Most of the people today are familiar divesting effects of HIV *i.e.*, the human immunodeficiency virus. The virus, which is transmitted by blood to blood contact, may produce no symptoms for years but typically within 10-15 years destroy the T_4-lymphocytes, the cell that play a key role on the immune system and causes a fatal disease *i.e.*, AIDS (acquired immunodeficiency syndrome)[1]. The resulting depletion in the level of essential immuno cells leave patients vulnerable to opportunistic infection that would not normally harm a healthy person[2].

Reverse transcriptase is the key enzyme of HIV, catalyzing the RNA-depending and DNA-dependant synthesis of double strand viral DNA[3]. HIV-1 reverse transcriptase is an attractive target for the drug therapy of AIDS, because it is essential for HIV replication and it is not required for normal host cell replication[4].

One class of reverse transcriptase inhibitors is nucleoside analogue (NRTIs) like AZT and DDI[5]. These dideoxy compounds cause DNA chain termination when they are incorporated in to a growing DNA strand. However, it is found that the treatment of some of these nucleoside inhibitors such as AZT is some time associated with bone marrow suppression[6]. Another class of HIV- reverse transcriptase inhibitors is non-nucleoside inhibitors (NNRTIs), which like the nucleoside analogues block reverse transcriptase but have a different mode of inhibition for viral replication. These inhibitors include TIBO, HEPT, BHAP and R-APA *etc*. Among them HEPT has proved to be a potent and selective inhibitor of HIV-1. Other animal retrovirus and even HIV-2 are totally unaffected by this compound[5].

EXPERIMENTAL

Data set: In the present work, structural information and corresponding biological activity data are taken from the literature[6]. The descriptors are calculated with the help of Hansch well characterized aliphatic substituents table[7]. The numeral value of this data was presented in Table-1. The functionalities of descriptors used to represent different structural modification by R_1 and R_2 was also tabulated in Table-2.

Statistical methods: Developing a QSAR model requires a diverse set of data, thereby a large number of descriptors have to be considered. Descriptors are numerical values that encode different structural features of the molecules. Selection of a set of appropriate descriptors from large number of them requires a method, which is able to discriminate between the parameter. Pearson correlation matrix has been performed on all descriptors by using NCSS statistical Software[8], shown in Table-3. The analysis of matrix revealed nine descriptors for the development of MLR model. The value of descriptors selected for MLR model are presented in Table-1. The model was than formed by a stepwise addition of terms. A deletion process was than employed, whereby each variable in the model was held out in turn and using the remaining parameters models were generated. Each descriptor was chosen as input for the statistical software package and then stepwise addition

TABLE-1
STRUCTURAL MODIFICATION, NNRTIS DATA OF BIOLOGICAL ACTIVITY [BA (obs.)] FOR HEPT ANALOGUES RELATED TO STUDY, DESCRIPTORS VALUE CALCULATED WITH THE HELP 'HANSCH, WELL CHARACTERIZED ALIPHATIC SUBSTITUENT TABLE, BIOLOGICAL ACTIVITY (calcd.); BIOLOGICAL ACTIVITY CALCULATED BY OF EQN. 1 AND PREDICTION ERROR; DIFFERENCE OF BA (obs.) AND BA (calcd.)

S. No.	R_1	R_2	BA (obs.)	R_1F_r	R_1H_{ACC}	R_1H_{DON}	R_1MR	R1 ζ	R_2F_r	R2MR	$R_2\zeta$	α/β#	BA (calcd.)	Prediction error
1	CH_2OH	Me	6.253	-1.1	1	1	7.1	0	0.77	5.65	-0.04	1	6.238	0.015
2	CH_2OH	Et	6.11	-1.1	1	1	7.1	0	1.43	10.3	-0.05	1	6.238	-0.128
3*	CH_2OH	n-Pr	5.204	-1.1	1	1	7.1	0	1.97	14.96	-0.06	1	6.238	-1.034
4	CH_2OH	i-Pr	6.955	-1.1	1	1	7.1	0	1.84	14.96	-0.05	1	6.238	0.717
5	H	Me	5.605	0	0	0	0	0	0.77	5.65	-0.04	1	6.590	-0.985
6*	H	Et	6.306	0	0	0	0	0	1.43	10.3	-0.05	1	6.590	-0.284
7	H	i-Pr	6.886	0	0	0	0	0	1.84	14.96	-0.05	1	6.590	0.296
8	CH_3	Me	6.777	0.77	0	0	5.65	-0.04	0.77	5.65	-0.04	1	6.836	-0.059
9*	CH_3	Et	6.983	0.77	0	0	5.65	-0.04	1.43	10.3	-0.05	1	6.836	0.147
10	CH_3	n-Pr	6.561	0.77	0	0	5.65	-0.04	1.97	14.96	-0.06	1	6.836	-0.275
11	CH_3	i-Pr	7.222	0.77	0	0	5.65	-0.04	1.84	14.96	-0.05	1	6.836	0.386
12*	C_6H_5	Me	7.125	1.2	0	0	25.1	0.08	0.77	5.65	-0.04	1	6.974	0.151
13	C_6H_5	Et	7.377	1.2	0	0	25.1	0.08	1.43	10.3	-0.05	1	6.974	0.403
14	C_6H_5	n-Pr	6.658	1.2	0	0	25.1	0.08	1.97	14.96	-0.06	1	6.974	-0.316
15*	C_6H_5	i-Pr	7.319	1.2	0	0	25.1	0.08	1.84	14.96	-0.05	1	6.974	0.345
16	CH_2OH	Me	4.231	-1.1	1	1	7.1	0	0.77	5.65	-0.04	-1	4.336	-0.105
17	CH_2OH	Et	4.248	-1.1	1	1	7.1	0	1.43	10.3	-0.05	-1	4.336	-0.088
18*	H	Me	4.588	0	0	0	0	0	0.77	5.65	-0.04	-1	4.688	-0.100
19	H	Et	4.909	0	0	0	0	0	1.43	10.3	-0.05	-1	4.688	0.221
20	H	n-Pr	4.228	0	0	0	0	0	1.97	14.96	-0.06	-1	4.688	-0.460
21*	CH_3	Me	5.355	0.77	0	0	5.65	-0.04	0.77	5.65	-0.04	-1	4.934	0.421
22	CH_3	Et	5.291	0.77	0	0	5.65	-0.04	1.43	10.3	-0.05	-1	4.934	0.357
23	C_6H_5	Me	5.163	1.2	0	0	25.1	0.08	0.77	5.65	-0.04	-1	5.072	0.091
24*	C_6H_5	Et	5.478	1.2	0	0	25.1	0.08	1.43	10.3	-0.05	-1	5.072	0.406
25	C_6H_5	n-Pr	5.015	1.2	0	0	25.1	0.08	1.97	14.96	-0.06	-1	5.072	-0.057

*S. No. are of test set molecule and remaining are of training set molecule. # α form was described by 1 and β form by -1.

TABLE-2
FUNCTIONALITIES OF DESCRIPTORS USED AS PHYSICOCHEMICAL PARAMETER IN STUDY

Functional families of descriptors	Descriptor definition
Constitutional descriptors	Hydrogen acceptor (H_{ACC}), hydrogen donor (H_{DON}), α - form, β - form
Steric Descriptors	Molecular refractivity (MR)
Hydrophobic Parameter	Substituent constant (F_r)
Electronic Parameter	Swain and Lupton field parameter (ζ)

method implemented in the software was used for choosing the descriptors contributing to the anti HIV activity of HEPT analogues. In fact, this was the study to harmonize α and β form, so α/β was the permanent descriptor for every study. This was supported by Table-3, which shows good correlation with biological activity and poor inter-correlation with other descriptors.

The specifications for the best selected MLR models with least number of descriptors are shown in Table-4. It is well known that there are three well known components in any QSAR study *i.e.*, development of models, validation of models and utility of the developed model. Validation is a crucial aspect of any QSAR analysis. The statistical quality of resulting model, as given in Table-4, is determined by R^2, SSE and F[9-11]. It is noteworthy that all these equations were derived using the entire data set of compounds (n = 17) and no outliers were found. The F-value presented in Table-4 is found statistically significant at 99.9 % level, science al the calculated F-value are higher as compared to tabulated value [$F_{2,17\alpha0.001} = 11.78$).

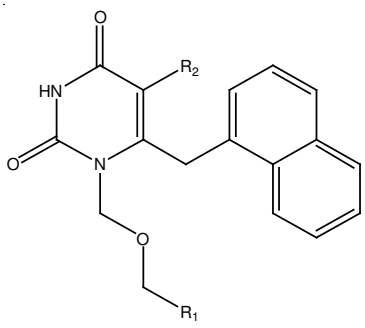
For testing the validity of predictive power of the selected MLR models the LOO technique was used. The developed model were validated by calculation of following statistical

TABLE-3
PEARSON CORRELATION MATRIX

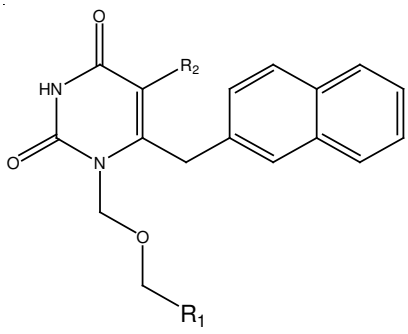
	BA	R_1Fr	R_1H_{ACC}	R_1H_{DON}	R_1MR	$R_1\zeta$	R_2Fr	R_2MR	$R_2\zeta$	α/β
BA	1	–	–	–	–	–	–	–	–	–
R_1Fr	0.275	1	–	–	–	–	–	–	–	–
R_1H_{ACC}	-0.179	-0.887	1	–	–	–	–	–	–	–
R_1H_{DON}	-0.179	-0.887	1	1	–	–	–	–	–	–
R_1MR	0.143	0.546	-0.157	-0.157	1	–	–	–	–	–
$R_1\zeta$	-0.044	0.367	-0.143	-0.143	0.856	1	–	–	–	–
R_2Fr	0.203	0.235	-0.254	-0.254	0.066	0.062	1	–	–	–
R_2MR	0.254	0.219	-0.247	-0.247	0.039	0.035	0.993	1	–	–
$R_2\zeta$	-0.022	0.318	0.304	0.304	-0.179	-0.171	-0.932	-0.889	1	–
α/β	0.891	0.001	0.015	0.015	-0.063	-0.153	0.070	0.118	0.068	1

TABLE-4
EQUATIONS GENERATED BY MLR ANALYSIS ALONG WITH PERTINENT STATISTICAL PARAMETER

Equations		Statistical parameter
$BA = 0.320(\pm 0.113)R_1Fr + 0.951(\pm 0.104)\alpha/\beta + 5.639 (\pm 0.105)$	(1)	$N = 17, R^2 = 0.869, SSE = 0.422, F = 46.285, R^2_{adi} = 0.850,$ $PRESS = 3.445, SSY = 19.002, S_{PRESS} = 0.450$ and $R^2_{CV} = 0.550$
$BA = -0.447(\pm 0.225)R_1H_{ACC} + 0.960(\pm 0.118)\alpha/\beta + 5.814(\pm 0.140)$	(2)	$N = 17, R^2 = 0.831, SSE = 0.479, F = 34.402, R^2_{adi} = 0.807,$ $PRESS = 4.517, SSY = 19.002, S_{PRESS} = 0.515$ and $R^2_{CV}=0.485$
$BA = -0.447(\pm 0.225)R_1H_{DON} + 0.960(\pm 0.118)\alpha/\beta + 5.814 (\pm 0.140)$	(3)	$N = 17, R^2 = 0.831, SSE = 0.479, F = 34.402, R^2_{adi} = 0.807, PRESS$ $= 4.517, SSY = 19.002, S_{PRESS} = 0.515$ and $R^2_{CV} = 0.485$
$BA = 0.023(\pm 0.013)R_1MR + 0.970(\pm 0.117)\alpha/\beta + 5.466(\pm 0.167)$	(4)	$N = 17, R^2 = 0.834, SSE = 0.475, F = 35.077, R^2_{adi} = 0.810, PRESS$ $= 4.481, SSY = 19.002, S_{PRESS} = 0.513$ and $R^2_{CV} = 0.487$
$BA = 2.357(\pm 3.000)R_1\zeta + 0.972(\pm 0.129)\alpha/\beta + 5.658(\pm 0.131)$	(5)	$N = 17, R^2 = 0.803, SSE = 0.518, F = 28.450, R^2_{adi} = 0.774, PRESS$ $= 5.445, SSY = 19.002, S_{PRESS} = 0.566$ and $R^2_{CV} = 0.434$
$BA = 0.312(\pm 0.257)\ R_2Fr + 0.946(\pm 0.124)\alpha/\beta + 5.232(\pm 0.387)$	(6)	$N = 17, R^2 = 0.814, SSE = 0.503, F = 30.619, R^2_{adi} = 0.787, PRESS$ $= 5.307, SSY = 19.002, S_{PRESS} = 0.559$ and $R^2_{CV} = 0.441$
$BA = -0.447(\pm 0.225)\ R_2MR + 0.938(\pm 0.124)\alpha/\beta + 5.224(\pm 0.361)$	(7)	$N = 17, R^2 = 0.816, SSE = 0.500, F = 31.060, R^2_{adi} = 0.790, PRESS$ $= 5.225, SSY = 19.002, S_{PRESS} = 0.554$ and $R^2_{CV} = 0.446$
$BA = -12.061(\pm 17.436)\ R_2\zeta + 0.963(\pm 0.128)\alpha/\beta + 5.086(\pm 0.873)$	(8)	$N = 17, R^2 = 0.801, SSE = 0.520, F = 28.113, R^2_{adi} = 0.772, PRESS$ $= 5.565, SSY = 19.002, S_{PRESS} = 0.572$ and $R^2_{CV} = 0.428$



α-form (1-15)



β-form (16-25)

Chemical structure of α and β-form for 1-alkoxymethyl-5-alkyl-6-naphthylmethyl uracils

parameters: PRESS, SSY, S_{PRESS}, R^2_{CV} and R^2_{adj} (Table-4). These parameters were calculated from following equations:

$$PRESS = \Sigma(Y_{obs} - Y_{calcd})^2$$
$$SSY = \Sigma(Y_{obs} - Y_{mean})^2$$
$$S_{PRESS} = \sqrt{(PRESS/n)}$$
$$R^2_{CV} = 1 - (PRESS/SSY)$$
$$R^2_{adj} = 1 - (R^2)\ [(n - 1)/(n - p - 1)]$$

where Y_{obs}, $Y_{calcd.}$ and Y_{mean} are observed, calculated and mean values; n = number of compounds; p = number of independent parameters.

PRESS is an acronym for prediction sum of squares. It is used to validate a regression model with regards to predictability. To calculate PRESS, each observation is individually omitted. The remaining n - 1 observations are used to calculate a regression and estimate the value of the omitted observation. This is done n times, once for each observation. The difference between the actual Y value, Y_{obs} and the predicted Y, $Y_{calcd.}$, is called the prediction error. The sum of the squared prediction errors is the PRESS value. The smaller PRESS explains the better predictability of the model. Its value being less than SSY points out that the model predicts better chance and can be considered statistically significant. SSY are the sums of squares associated with the corresponding source of variation. These values are in terms of the dependent variable, Y.

The PRESS value above can be used to compute an R^2_{CV} statistic, called R^2 cross validated, which reflect the prediction ability of the model. This is a good way to validate the prediction of a regression model without selecting another sample or splitting present data. It is possible to have a high R^2 value and low R^2_{CV}. When this occurs, it is implies that the fitted model is data dependent. This R^2_{CV} ranges from below zero to above one. When outside range from zero to one, it is truncated to say with in this range. Adjusted R-square is an adjusted version of R^2. The adjustment seeks to remove the distortion due to a small sample size.

In many cases R^2_{CV} and R^2_{adj} is taken as a proof of the high ability of QSAR models. A high value of these statistical characteristic (> 0.5) is considered as a proof of high predictive ability of the model, although recent reports have proven the poosite[12]. Although a low value of R^2_{CV} for the training set can be indeed serving as an indicator of a low predictive ability of a model, the opposite is not necessarily true. Indeed, the high R^2_{CV} does not imply automatically high predictive ability of the model. Thus the high value of LOO R^2_{CV} is necessary condition for a model to have a high predictive power; it is not a sufficient condition. It is proven that only to estimate the true predictive power of a model is to test it on a sufficiently large collection of compounds from an external test set. This application is necessary for obtaining trustful statistics for the comparison between the observed and predicted activities these compounds. Beside high R^2_{CV}, a reliable model should be also characterized by a high correlation coefficient between the predicted and observed activities of the compounds from a test set of molecules that was not used to develop the model.

On behalf of above discussion, eqn. 1 gave best statistical agreement among the 8 equations, so considered as best model to conclude the QSAR prediction. To confirm the predictive power of QSAR models, an external set of HEPT analogues was used. Eight corresponding HEPT analogues were used as external set of molecules.

RESULTS AND DISCUSSION

Exercise and output of training set: In the present study we tried to develop best model to explain the correlation between the physicochemical parameter and HIV non-nucleoside reverse transcriptase inhibitor activities of HEPT analogues. Among the best QSAR equation, the best QSAR model were selected on the basis various statistical parameter such as squared correlation coefficient ($R^2 > 0.64$), root mean standard error of estimate (SSE < 0.5) and F-test value at 99.9 % significance level with least number of descriptors included in equation. Some of best equations are presented in Table-4. The generated best model was validated for predictive ability inside the model (leave one out method) and outside model (test and training set).

Among these equations, the eqn. 1 was considered the best model explaining 86.9 % variance in activity. The low standard error of estimate(s), high F value and other statistical parameter suggest that the model is statistically highly significant. The data showed overall statistical significance > 99.9 % with F = 11.78. The graphical presentation BA (obs.) and BA (calcd.) with eqn. 1 was shown as Fig. 1 for training set molecule, where the squared correlation coefficient value (R^2 = 0.867) was observed.

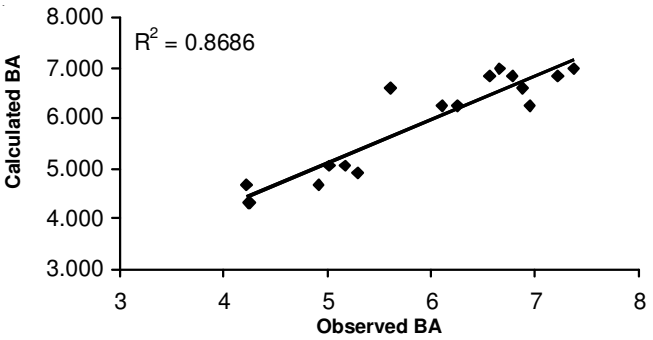


Fig. 1. Graphical presentation biological activity (BA) (obs.) and biological activity (calcd.) with for training set molecules

External validation: The validation of the best model (eqn. 1) has been done on a test set of 8 compounds, where good squared correlation coefficient ($r^2 = 0.779$) was observed between the predicted and observed activity. The graphical presentation biological activity (obs.) and biological activity (calcd.) with eqn. 1 was shown in Fig. 2, for test set molecule.

Conclusion

2D QSAR model have been developed to estimate and predict anti HIV activities against HIV-1 for α and β forms of HEPT analogues (1-alkoxymethyl-5-alkyl-6-naphthylmethyl uracils). First, 15 HEPT analogues of α-form were studied, again 10 HEPT analogues of β form were studied, both studies

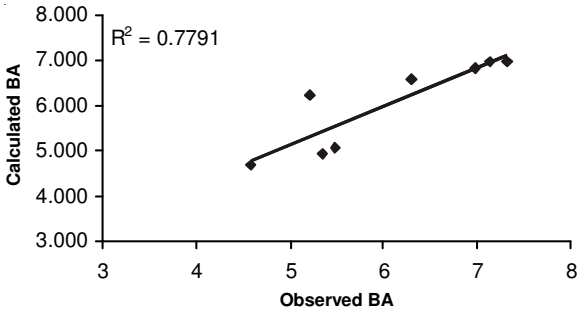


Fig. 2. Graphical presentation biological activity (BA) (obs.) and biological activity (calcd.) with for test set molecules

was having good statistical significance, separately. But, for better understanding of the model, both the studies are harmonized. The conclusion for harmonised study was $R^2 = 0.869$ up on a series for 17 compounds as HEPT ligands for training set molecule, inhibitors of HIV reverse transcriptase; using the QSAR that imply analysis of correlation and multi linear regression; a significant collection of descriptors was used. Eqn. 1 gave good squared correlation coefficient ($r^2 = 0.779$) with test set molecules.

Hence QSAR studies provide deeper insight into the mechanism of action of compounds that ultimately becomes of great importance in modification of the structure of compounds. In addition, QSAR quantitative models which permit prediction of activity of compounds prior to synthesis.

This will help in rationalizing the design of novel and potent analogues. From the descriptors incorporated in the QSAR model, one may conjecture that increase the hydrophobicity at R_1 will enhance the inhibitory activity.

Therefore, this QSAR study on the series of 1-alkoxymethyl-5-alkyl-6-naphthylmethyl uracils as HEPT analogues strongly support the study published by Aarei and Atabati[6]. Infect, this study done by using classical QSAR (different approach), comparatively very less expensive technique and providing approximately same quality conclusion as previously published study.

REFERENCES

1. S.J. O'Brien and M. Dean, *Sci. Am.*, **227**, 44 (1997).
2. S Mishra, S.P. Dwivedi, N. Dwivedi and R.B. Singh, *The Open Nutraceut. J.*, **2**, 46 (2009).
3. D.B. Kireev, J.R. Chretien, D.S. Grierson and C. Monneret, *J. Med. Chem.*, **40**, 4257 (1997).
4. S. Hannongbua, K. Nivesanond, L. Lawtrakul, P. Pungpo and P. Wolschann, *J. Chem. Inf. Comp. Sci.*, **41**, 848 (2001).
5. R. Garg, S.P. Gupta, H. Gao, M.S. Babu, A.K. Debnath and C. Hansch, *Chem. Rev.*, **99**, 3525 (1999).
6. K. Aarei and M. Atabati, *J. Chin. Chem. Soc.*, **56**, 206 (2009).
7. C. Hansch and A. Leo, Substituents Constant for Correlation Analysis: In Chemistry and Biology, ACS Pub. (1979).
8. NCSS Statistical Software, Available online: http://www.ncss.com, Accessed Aug 27 (2010).
9. G.W. Snedecor and W.G. Cochran, Statistical Methods; Oxford and IBH: New Delhi, India, p. 381 (1967).
10. S. Chatterjee, A.S. Hadi and B. Price, Regression Analysis by Examples; Wiley VCH: NewYork, USA (2000).
11. M.V. Diudea, QSPR/QSAR Studies for Molecular Descriptors; Nova Science: Huntingdon, New York, USA (2000).
12. A. Golbraikh and J. Tropha, *J. Mol. Graph. Mod.*, **20**, 269 (2002).