Molecular Diversity (2006) 10: 405–414
DOI: 10.1007/s11030-005-9012-2

Full-length paper

A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis

Antreas Afantitis[1,4], Georgia Melagraki[1], Haralambos Sarimveis[1,*], Panayiotis A. Koutentis[2], John Markopoulos[3] & Olga Igglessi-Markopoulou[1]
[1]*School of Chemical Engineering, National Technical University of Athens, Athens, Greece;* [2]*Department of Chemistry, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus;* [3]*Department of Chemistry, University of Athens, Athens, Greece;* [4]*Department of ChemoInformatics, NovaMechanics Ltd, Cyprus*
*(*Author for correspondence, E-mail: hsarimv@central.ntua.gr, Tel.: +30-210-7723237, Fax: +30-210-7723138)*

Received 24 June 2005; Accepted 1 November 2005

Key words: anti-HIV, HEPT derivatives, stepwise regression, MLR, QSAR

Summary

A quantitative–structure activity relationship was obtained by applying Multiple Linear Regression Analysis to a series of 80 1-[2-hydroxyethoxy-methyl]-6-(phenylthio) thymine (HEPT) derivatives with significant anti-HIV activity. For the selection of the best among 37 different descriptors, the Elimination Selection Stepwise Regression Method (ES-SWR) was utilized. The resulting QSAR model ($R^2_{CV} = 0.8160$; $S_{PRESS} = 0.5680$) proved to be very accurate both in training and predictive stages.

Introduction

Novel medicines are typically developed using a trial-and-error approach which is costly and time-consuming. The application of quantitative–structure activity relationship (QSAR) methodologies to this problem has the potential to greatly decrease the time and effort required to improve current medicines in terms of their efficacy or to discover new ones. A successful QSAR model generates statistically significant relationships between chemical structure and biological activity [1].

Human immunodeficiency virus type 1 (HIV-1) is the primary cause of AIDS (acquired immunodeficiency syndrome), which is one of the main medical and social problems in our epoch. *HEPT* 1-(2-hydroxyethoxy-methyl)-6-(phenylthio) thymine derivatives (Scheme 1) are the first non nucleoside reverse transcriptase inhibitor (NNRTI) analogues shown to have both potent anti-HIV activity and inhibit HIV-1 at nanomolar concentration [2]. The design of new HEPT derivatives requires a more detailed knowledge of the mechanism of reverse transcriptase (RT) inhibition by this class of compounds. QSAR is a powerful approach to discern chemical properties of compounds that are required for specific biological activity [3, 4].

In the past, numerous attempts have been made to predict the molar concentration of a drug required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 [5]. Luco et al. [6] used multiple linear regression (MLR) and partial least squares (PLS) methods but no external data sets were used to validate the models. Subsequently Jalali-Heravi and Parastar [7] used Luco's MLR model in order to test it with external data and furthermore they introduced a new MLR model and a new non-linear model based on artificial neural networks (ANN). Using Principal Component Analysis (PCA) and 36 derivatives, Alves et al. [8] presented a discussion for the significance of each variable to anti-HIV activity. Bazoui et al. [9] used MLR and ANN computational techniques but no external compounds were used to validate the models. Douali et al. [10–12] introduced a QSAR model using ANNs and a part of Luco's data set. The predictive ability of the models was tested with the use of leave-one-out (LOO) procedure. Gupta et al. [13] predicted the anti-HIV activity using an eccentric adjacency index. Using a different data set from Luco et al. [6], Gayen et al. [14] presented very good QSAR models based on the MLR technique.

In this work we used a data set of 80 HEPT derivatives [6, 7, 10] that has been reported in the literature as reliable. 37 topological and structural descriptors were considered. Among them, the most statistically significant descriptors were selected, using a rigorous variable selection method. The result of this study was the development a new linear QSARs model containing 5 variables. In order to validate the proposed methodology, we used two validation strategies:

Table 1. Topological and structural descriptors

Descriptor ID	Description	Notation		Description	Notation
1	Reciprocal of the standard shadow area shadow on YZ plane	1/S [7]	2	Connectivity index	${}^4\chi_p^N$ [6]
3	Ratio of the partial charges on the most positive and the most negative atoms	POS/NEG [7]	4	Molecular Volume	V_x [6]
5	Heat of Formation	ΔH_f (kcal/mol) [7]	6	Verloop steric parameter	B1-3R_1 [6]
7	Square of the number of SP^3 carbon atoms on the R_2 substituent	(NCSP^3-R_2)^2 [7]	8	Taft steric constant for ortho substituents	Es-2R_1[6]
9	Cub of summation of the positions of R_1 on the C-6 aromatic ring constant	(NS-R_1)^3 [7]	10	Connectivity index	${}^\circ\Delta\chi(R_3)$ [6]
11	Number of hydroxyl groups on the R_3 substituent	NOH-R_3 [7]	12	Connectivity (chain) index	${}^6\chi_{ch}^v$ [6]
13	Hansch constant	$\Sigma\pi$ ($\pi R_1+\pi R_2$) [6]	14	Indicator parameter (takes the value 1 or 0 for the presence or absence of a six membered saturated ring in R_3)	Ich-R_3 [6]
15	Connectivity level (dividing the value of the ${}^1\chi(R_2)$ index by the number of atoms involved in their calculus)	${}^1\chi^N(R_2)$ [6]	16	Indicator parameter (takes the value 1 or 0 for the presence or absence of a substituent at 4 – position of the C–6 aromatic ring	I-4R_1 [6]
17	Molar Refractivity	MR	18	Diameter	Diam
19	Partition Coefficient (Octanol Water)	ClogP	20	Molecular Topological Index	TIndx
21	Principal Moment of Inertia Z	PMIZ	22	Number of Rotatable Bonds	NRBo
23	Principal Moment of Inertia Y	PMIY	24	Polar Surface Area	PSAr
25	Principal Moment of Inertia X	PMIX	26	Radius	Rad
27	Connolly Accessible Area	SAS	28	Shape attribute	ShpA
29	Total Energy	TotE	30	Sum of Valence Degrees	SVDe
31	LUMO Energy	LUMO	32	Total Connectivity	TCon
33	HOMO Energy	HOMO	34	Total Valence Connectivity	TVCon
35	Balaban Index	BIndx	36	Wiener Index	WIndx
37	Cluster Count	ClsC			

Y-randomization and external validation using division of the entire dataset set into training and test sets.

Materials and methods

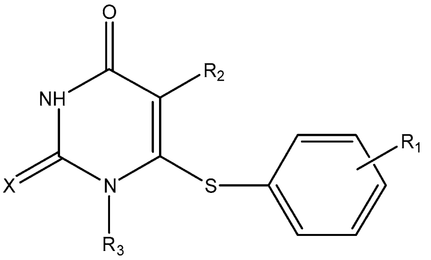
Data set

In this QSAR study 80 of the 107 HEPT derivatives of the Luco and Fereti [6] data set were used. The biological activities of these 80 compounds were reported in the same paper [6]. In order to model and predict the specific activity (the molar concentration of a drug required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1), 37 physicochemical constants, topological and structural descriptors (Table 1) were considered as possible input candidates to the model. The first 16 descriptors were collected from the literature [6, 7] and the rest of them were calculated with ChemSar which is included in Chemoffice.

The objective of this work was to determinate a subset of variables which afford the most significant linear QSAR models linking the structure of these compounds with their anti-HIV activity.

Stepwise multiple regression

As mentioned in the introduction, the ES-SWR algorithm [15] was used to select the most appropriate descriptors. ES-SWR is a popular stepwise technique which combines Forward Selection (FS-SWR) and Backward Elimination (BE-SWR). It is basically a forward selection approach, but at each step it considers the possibility of deleting a variable as in the backward elimination approach, provided that the number of model variables is greater than two. The two basic elements of the ES-SWR method are described next in more details.



Scheme 1. HEPT Derivatives

Forward selection

The variable considered for inclusion at any step is the one yielding the largest single degree of freedom F-ratio among the variables that are eligible for inclusion. The variable is included only if this value is larger than a fixed value F_{in}. Consequently, at each step, the jth variable is added to a k-size model if

$$F_j = \max_j \left(\frac{RSS_k - RSS_{k+j}}{s^2_{k+j}} \right) > F_{in} \tag{1}$$

In the above inequality *RSS* is the *residual sum of squares* and *s* is the *mean square error.* The subscript *k*+*j* refers to quantities computed when the *j*th variable is added to the *k* variables that are already included in the model.

Backward elimination

The variable considered for elimination at any step is the one yielding the minimum single degree of freedom *F*-ratio among the variables that are included in the model. The variable is eliminated only if this value does not exceed a specified value $F_{out.}$ Consequently, at each step, the *j*th variable is eliminated from a *k*-size model if

$$F_j = \min_j \left(\frac{RSS_{k-j} - RSS_k}{s^2_k} \right) < F_{out} \tag{2}$$

The subscript $k - j$ refers to quantities computed when the *j*th variable is eliminated from the *k* variables that have been included in the model so far.

Cross-validation technique

In order to explore the reliability of the proposed method we also used the cross-validation method. Based on the cross-validation technique, a number of modified data sets are created by deleting in each case one (LOO) or a small group (leave-some-out) of objects [16]. For each data set, an input-output model is developed, based on the utilized modelling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). In particular, the LOO procedure was utilized in this study, which produces a number of models, by deleting each time one object from the training set. Obviously, the number of models produced by the LOO procedure is equal to the number of available examples *n*. Prediction error sum of squares (*PRESS*) is a standard index to measure the accuracy of a modelling method based on the cross-validation technique. Based on the *PRESS* and *SSY* (Sum of squares of deviations of the experimental values from their mean) statistics, the R^2_{CV} and S_{PRESS} values can be easily calculated. The formulae used to calculate all the aforementioned statistics are presented below (Equations 3 and 4):

$$R^2_{CV} = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n} (y_{\exp} - y_{pred})^2}{\sum_{i=1}^{n} (y_{\exp} - \bar{y})^2} \tag{3}$$

$$S_{PRESS} = \sqrt{\frac{PRESS}{n}} \tag{4}$$

Y- randomization test

This technique ensures the robustness of a QSAR model [17, 18]. The dependent variable vector (biological action) is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to have low R^2 and R^2_{CV} values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

Estimation of the predictive ability of a QSAR model

According to Tropsha et al. [18] the predictive power of a QSAR model can be conveniently estimated by an external R^2_{CVext} (Eq. 5).

$$R^2_{CVext} = 1 - \frac{\sum_{i=1}^{test} (y_{\exp} - y_{pred})^2}{\sum_{i=1}^{test} (y_{\exp} - \bar{y}_{tr})^2} \tag{5}$$

where $\bar{y}_{tr}$ is the averaged value for the dependent variable for the training set.

Furthermore Tropsha et al. [18, 19] considered a QSAR model predictive, if the following conditions are satisfied:

$$R^2_{CVext} > 0.5 \tag{6}$$

$$R^2 > 0.6 \tag{7}$$

$$\frac{(R^2 - R^2_{\text{o}})}{R^2} < 0.1 \quad \text{or} \quad \frac{(R^2 - R'^2_{\text{o}})}{R^2} < 0.1 \tag{8}$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \tag{9}$$

Mathematical definitions of R^2_{o}, $R^{'2}_{\text{o}}$, k and k' are based on regression of the observed activities against predicted activities and the opposite (regression of the predicted activities against observed activities). The definitions are presented clearly in ref. 19 and are not repeated here for brevity.

Defining model applicability domain

In order for a QSAR model to be used for screening new compounds, its domain of application [18, 20] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of Extrapolation* [18] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [21] for each chemical, where the QSAR model is used to predict its activity:

$$h_i = x_i^T (X^T X) x_i \tag{10}$$

In Equation (10) x_i is the descriptor-row vector of the query compound and *X* is the $k \times n$ matrix containing the *k* descriptor values for each one of the *n* training compounds. A leverage value greater than $3k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may not reliable.

Results and discussion

For the selection of the most important descriptors, the aforementioned stepwise multiple regression technique was used. In order to automate the above procedure, we developed in-house a software that realizes the ES-SWR algorithm. More specifically the algorithm was programmed in the MATLAB programming language and is quite generic, so that it can accept a practically unlimited number of descriptors.

The produced MLR models were studied thoroughly as new descriptors were selected by the aforementioned variable selection method. The most significant descriptor according to the ES-SWR algorithm is the connectivity index ${}^4\chi_p^N$ followed by the connectivity (chain) index ${}^6\chi_{ch}^v$, and the number of hydroxyl groups on the R_3 substituent NOH-R_3. The connectivity index ${}^4\chi_p^N$ describes the shape characteristics of the entire module, which according to Luco et al. [6] is the most important variable. The connectivity (chain) index ${}^6\chi_{ch}^v$ encodes information about the number and type of six-membered rings present in the module. NOH-R_3 is a simple topological descriptor and indicates the number of hydroxyl groups on the R_3 substituent. The MLR model that consists of only the three most significant descriptors is already quite accurate and attractive as well, since all three descriptors are topological indices, they are calculated rapidly and have a clear physical meaning.

However, the performance of the MLR model was improved substantially, by including the next two most significant descriptors: HOMO energy and Lipophilicity (ClogP). Molecular orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to Frontier Orbital Theory, the shapes and symmetries of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) are crucial in predicting the reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction. All the structures before the calculation of the HOMO Energy were fully optimized using the AM1 basis set. Lipophilicity is known to be important for absorption, permeability, and *in vivo* distribution of organic compounds [21, 22] and has been used as a physicochemical descriptor in QSARs with great success [1, 24].

Adding more descriptors to the 5-parameter model, did not improve significantly the prediction abilities of the models, in terms of R^2_{CV}. For brevity, the remaining analysis will be focused on the 5-parameter model. Nevertheless, the rest of the models can be available to the interested readers. In order to avoid internal correlations, we performed a correlation analysis on the five selected descriptors and the results are presented in Table 2. All the values deviate from unity considerably so there is no significant correlation between the five independent variables.

Table 2. Correlation matrix of the 5 selected descriptors

	${}^4\chi_p^N$	${}^6\chi_{ch}^v$	NOH-R_3	HOMO	CLogP
${}^4\chi_p^N$	1,00				
${}^6\chi_{ch}^v$	0,428	1,00			
NOH-R_3	−0,336	−0,373	1,00		
HOMO	0,190	−0,103	0,098	1,00	
CLogP	0,754	0,612	−0,606	0,028	1,00

The full linear equation for the prediction of the anti-HIV activity $\log 1/c$ is the following:

$$\begin{aligned}\log\tfrac{1}{c} &= 66.30\,{}^4\chi_p^N - 28.80\,{}^6\chi_{ch}^v - 5.25{*}10^{-1}\text{NOH-R}_3\\ &\quad + 5.66{*}10^{-1}\text{HOMO} + 3.49 \ast 10^{-1}\text{CLogP} - 6.98\\ R^2 &= 0.841\ F = 77.99\ RMSE = 0.531\ R^2_{CV} = 0.816\\ S_{PRESS} &= 0, 568n = 80\end{aligned} \tag{11}$$

In order to further explore the prediction ability of the selected descriptors, the data set of 80 1-[2-hydroxyethoxy-methyl]-6-(phenylthio)thymine (HEPT) derivatives was divided into a training set of 60 compounds, and a validation set of 20 compounds. The selection of the derivatives in the training set was made according to the structure and the scale of the biological action, so that representatives of a wide range of structures (in terms of the different substituents, atoms and action) were included. According to Golbraikh and Tropsha [25] this approach is correct since representative points of the test set must be close to those of training set and vice versa. The compounds that constituted the training and validation sets are clearly presented in Table 3. The validation examples are marked with ^{a}.

Using the five descriptors, we developed a new MLR equation based on only the 60 training examples.

$$\begin{aligned}\log\tfrac{1}{c} &= 67.10\,{}^4\chi_p^N - 25.40\,{}^6\chi_{ch}^v - 5.29{*}10^{-1}\text{NOH-R}_3\\ &\quad + 6.18{*}10^{-1}\text{HOMO} + 3.66{*}10^{-1}\text{CLogP} - 6.81\\ R^2 &= 0.838\ F = 55.99\ RMSE\ = 0.560\ R^2_{CV} = 0.804\\ S_{PRESS} &= 0.616\ n = 60\end{aligned} \tag{12}$$

This equation was used to predict the HIV-activity for the validation examples. The results are presented in the last column of Table 3 and correspond to an R^2 value of 0.904. The residual plot for the predicting set is presented in Figure 1. The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure are adequate to generate an efficient QSAR model for predicting the HIV-activity of different compounds.

Table 3. Training and test set

A/A	R^1	R^2	R^3	X	Experimental Values	Predicted Values	
1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	4.999	
2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	4.985	
3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	5.327	
4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	5.201	
5[a]	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	–	5.234
6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	4.822	
7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	4.435	
8[a]	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	–	4.973
9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	5.187	
10[a]	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	–	5.236
11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.326	
12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.179	
13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	4.720	
14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	4.804	
15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	6.271	
16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	6.235	
17[a]	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	–	6.295
18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	4.838	
19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	4.670	
20[a]	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	–	4.782
21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.315	
22[a]	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	–	5.742
23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	5.488	
24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.23	6.891	
25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	7.261	
26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.379	
27[a]	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	–	7.235
28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92	5.719	
29[a]	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47		5.465
30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.20	6.863	
31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	7.237	
32	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.351	
33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	8.088	
34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66	4.588	
35[a]	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	–	4.632
36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	4.654	
37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	4.822	
38	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	5.648	
39[a]	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.22	–	5.553
40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	5.629	
41	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	6.07	6.386	
42	H	Me	$CH_2OCH_2CH_2Me$	O	5.06	5.114	
43[a]	H	Me	$CH_2OCH_2CH_2Ac$	O	5.17	–	4.491
44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12	5.754	
45	H	Me	CH_2OCH_2Me	O	6.48	5.876	
46	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.528	
47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	–	5.279
48	H	Me	$CH_2OCH_2CH_2F$	O	5.96	5.476	
49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.822	
50	H	Me	CH_2OCH_2Ph	O	7.06	–	6.522
51	H	Et	CH_2OCH_2Me	O	7.72	6.828	
52	H	Et	CH_2OCH_2Me	S	7.58	6.848	
53	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	–	8.366
54	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.387	
55	H	Et	CH_2OCH_2Ph	O	8.23	7.427	
56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	8.740	
57	H	Et	CH_2OCH_2Ph	S	8.09	–	7.453
58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	8.766	

(*Continued to next page*)

Table 3. Continued

A/A	R^1	R^2	R^3	X	Experimental Values	Predicted Values	
59	H	i-Pr	CH_2OCH_2Me	O	7.99	8.117	
60	H	i-Pr	CH_2OCH_2Ph	O	8.51	–	8.426
61	H	i-Pr	CH_2OCH_2Me	S	7.89	8.143	
62	H	i-Pr	CH_2OCH_2Ph	S	8.14	8.455	
63	H	Me	CH_2OMe	O	5.68	6.278	
64	H	Me	CH_2OBu	O	5.33	5.755	
65	H	Me	Et	O	5.66	6.666	
66[a]	H	Me	Bu	O	5.92	–	5.748
67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	8.458	
68	H	Et	CH_2O-i-Pr	S	6.66	6.437	
69[a]	H	Et	CH_2O-c-Hex	S	5.79	–	5.755
70	H	Et	CH_2OCH_2-c-Hex	S	6.45	5.734	
71	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	7.094	
72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.118	
73[a]	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	–	6.985
74	3,5-Cl_2	Et	CH_2OCH_2Me	S	8.13	8.651	
75	H	Et	CH_2O-i-Pr	O	6.47	8.159	
76	H	Et	CH_2O-c-Hex	O	5.40	6.424	
78	H	Et	$CH_2OCH_2CH_2Ph$	O	7.02	5.664	
77[a]	H	Et	CH_2OCH_2-c-Hex	O	6.35	–	5.565
79	H	c-Pr	CH_2OCH_2Me	S	7.02	7.266	
80	H	c-Pr	CH_2OCH_2Me	O	7.00	7.075	

[a] The test set.

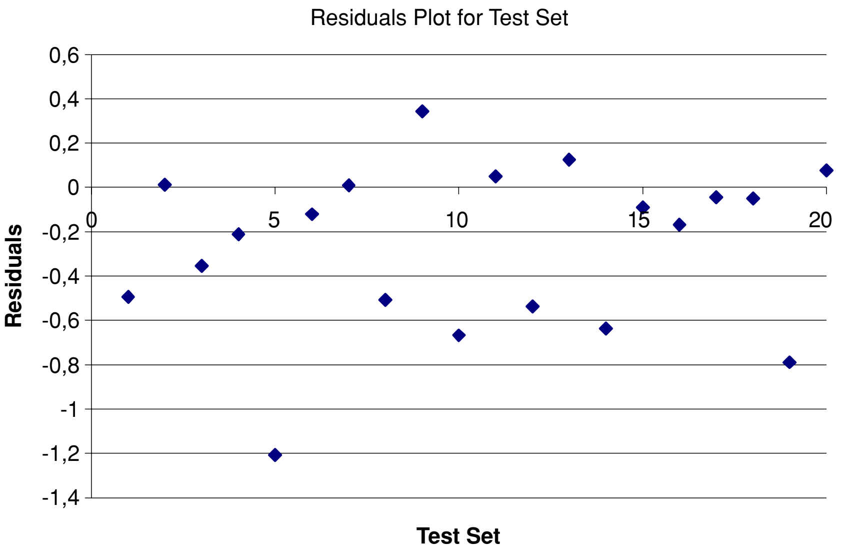


Figure 1. Residuals Plot for Test Set.

The proposed model (Eq. 12) passed all the tests for the predictive ability (Eqs. 6–9):

$$R^2_{CVext} = 0.827 > 0.5$$

$$R^2 = 0.904 > 0.6$$

$$\frac{\left(R^2 - R_0^2\right)}{R^2} = -0.0975 < 0.1$$

or

$$\frac{\left(R^2 - R_0'^2\right)}{R^2} = -0.0695 < 0.1$$

$$k = 1.0443 \text{ and } k' = 0.9543$$

The model was further validated by applying the Y-randomization. Several random shuffles of the Y vector were performed and the results are shown in Table 4. The low R^2 and R^2_{CV} values show that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

Finally the extent of extrapolation method was applied to the compounds that constitute the test set. The leverages for all 20 compounds were computed and are presented in Table 5. Two compounds (69 and 77) were found to fall slightly outside from the domain of the model (warning leverage limit 0.3). Both compounds contain cyclohexyl groups.

Table 4. Results of the Y-randomization test

Iteration	R^2	R^2_{CV}
1	0.073	0.00
2	0.045	0.00
3	0.039	0.00
4	0.072	0.00
5	0.047	0.00
6	0.085	0.00
7	0.103	0.00
8	0.073	0.00
9	0.078	0.00
10	0.194	0.00

Table 5. Leverages for the test set

Compound Id	Leverages
5	0.0554
8	0.0484
10	0.0493
17	0.0516
20	0.0686
22	0.0464
27	0.1491
29	0.0397
35	0.0506
39	0.0731
43	0.1548
47	0.1428
50	0.0521
53	0.0963
57	0.0526
60	0.1655
66	0.0624
69	0.3489
73	0.0639
77	0.3264

The proposed model, due to the high predictive ability, can therefore act as a useful aid to the costly and time consuming experiments for determining the molar concentration of a drug required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1. The model can also be used to screen existing databases or virtual libraries to identify new potentially active compounds. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" compounds.

We first tried to identify descriptors trends which lead to improved anti-HIV activity through modifications of the database compounds (Vanyur et al., 26) Based on the proposed QSAR equation we can make the following remarks regarding the importance of the descriptors and their effects on the anti HIV-activity: As mentioned before, our study agrees with the result of Luco et al. [6] that connectivity index ${}^4\chi_p^N$ is the most important variable for predicting the anti-HIV activity. As the value of descriptor ${}^4\chi_p^N$ increases so does the activity. An opposite effect on the activity is caused by an increase in the number of benzene rings (represented by ${}^6\chi_{ch}^v$), so special care should be taken when increasing the number of benzenes. Furthermore, the QSAR equation (Eq. 12) clearly shows that the presence of –OH group at the R_3 position decreases the activity. Finally, anti-HIV activity improves when lipophilicity (CLogP) or the HOMO energy take higher values. LogP and HOMO energy play a crucial role in distinguishing derivatives with X=O and those with X=S since all other three descriptors cannot recognize the difference between X=O and X=S.

An attempt was made to screen virtual libraries in order to identify novel potent compounds (Table 6). The introduction of a phenyl substituent at R_3 resulted in improved activities, although the leverages approached the limit (0.3) and in several cases exceeded the limit (>0.3) of the domain of applicability reducing the confidence of the predicted data. Introduction of an alkyl or alkoxy spacer between the phenyl substituent and the ring nitrogen at R_3 reduced the activities but improved the predictability. A good balance between activity and predictability was found with the introduction of a benzyl group at R_3. Furthermore the introduction of sulfur (X=S) in nearly all cases resulted in better activity than the oxygen (X=O) analogues. Exceptions to the latter were compounds where alkoxy chains were present at R_3, in these cases (A/A 23 to 38, Table 6) there was no significant difference in activities between the oxygen (X=O) and sulfur (X=S) species. R_2 also played an important role in increasing the activity, (see A/A 13 to 20, Table 6) and the bulky i-Pr group showed better activity than sterically less demanding n-Pr, Et and Me substituents. In this case the introduction of i-Pr also resulted in a reduced confidence in predictability since the limits of the domain of applicability were being approached as the molecular weight of the alkyl group increased. The effect of substitution at R_1 was more complex to analyse, however some observations can be made. Firstly mono- or disubstitution with methyl or halogen substituents was tolerated by the model whilst hydroxy or amino substituents were not; in particular the dihydroxy and diamino derivatives fell outside the domain of applicability (see A/A 73 to 76, Table 6). Secondly disubstitution generally gave better activities but at the expense of predictability. Finally no clear trend as to the preferred site of substitution was evident and may possibly be substituent dependant. Several interesting compounds (cf. A/A 20, 42 and 80, Table 6) were identified to have good activities (10.22, 10.48 and 10.15 respectively) and be comfortably within the domain of applicability and these are worthy of further study.

Conclusion

Our results lead to the conclusion that the anti-HIV activity of the HEPT derivatives can be successfully modelled with physicochemical constants and structural descriptors. The separation of the data into two independent sets (training

Table 6. Virtual screening results

A/A	R_1	R_2	R_3	X	log (1/IC_{50}) pred	Leverages (limit 0.30)
1	3,5-$(Cl)_2$	Et	Ph	O	9.5820	0.3968
2	3,5-$(Cl)_2$	Et	Ph	S	10.6013	0.2460
3	2-Me	Me	Ph	O	7.7403	0.1813
4	2-Me	Me	Ph	S	8.5856	0.1048
5	3-Me	Me	Ph	O	7.7632	0.2396
6	3-Me	Me	Ph	S	8.7188	0.1157
7	4-Me	Me	Ph	O	7.1742	0.1347
8	4-Me	Me	Ph	S	8.1109	0.0746
9	3,5 $(Me)_2$	Me	Ph	O	8.8133	0.3334
10	3,5 $(Me)_2$	Me	Ph	S	9.7612	0.1868
11	3,5 $(Me)_2$	Et	Ph	O	9.5406	0.3988
12	3,5 $(Me)_2$	Et	Ph	S	10.4847	0.2512
13	3,5 $(Me)_2$	Me	CH_2Ph	O	8.4658	0.1107
14	3,5 $(Me)_2$	Me	CH_2Ph	S	8.8250	0.1014
15	3,5 $(Me)_2$	Et	CH_2Ph	O	9.3378	0.1543
16	3,5 $(Me)_2$	Et	CH_2Ph	S	9.6970	0.1488
17	3,5 $(Me)_2$	Pr	CH_2Ph	O	8.9872	0.1095
18	3,5 $(Me)_2$	Pr	CH_2Ph	S	9.3470	0.1912
19	3,5 $(Me)_2$	i-Pr	CH_2Ph	O	9.8526	0.1931
20	3,5 $(Me)_2$	i-Pr	CH_2Ph	S	10.2207	0.2099
21	3,5 $(Me)_2$	i-Pr	CH_2CH_2Ph	O	9.8746	0.2174
22	3,5 $(Me)_2$	i-Pr	CH_2CH_2Ph	S	10.1117	0.2525
23	3,5 $(Me)_2$	i-Pr	OCH_2Ph	O	9.7923	0.1834
24	3,5 $(Me)_2$	i-Pr	OCH_2Ph	S	9.8080	0.1815
25	3,5 $(Me)_2$	i-Pr	OCH_2Me	O	10.0459	0.2063
26	3,5 $(Me)_2$	i-Pr	OCH_2Me	S	10.0639	0.1936
27	3,5 $(Me)_2$	i-Pr	CH_2OMe	O	10.0277	0.2449
28	3,5 $(Me)_2$	i-Pr	CH_2OMe	S	10.0539	0.2213
29	3,5 $(Me)_2$	i-Pr	CH_2OEt	O	9.6214	0.2026
30	3,5 $(Me)_2$	i-Pr	CH_2OEt	S	9.6490	0.1925
31	3,5 $(Me)_2$	i-Pr	CH_2OPh	O	9.8761	0.2040
32	3,5 $(Me)_2$	i-Pr	CH_2OPh	S	9.9068	0.2035
33	3,5 $(Me)_2$	i-Pr	$CH_2OCH_2CH_2OH$	O	8.7066	0.1655
34	3,5 $(Me)_2$	i-Pr	$CH_2OCH_2CH_2OH$	S	8.6856	0.1577
35	3,5 $(Me)_2$	i-Pr	$CH_2OCH_2CH_2OMe$	O	9.2237	0.1707
36	3,5 $(Me)_2$	i-Pr	$CH_2OCH_2CH_2OMe$	S	9.1986	0.1592
37	3,5 $(Me)_2$	Et	CH_2OEt	O	8.7727	0.1885
38	3,5 $(Me)_2$	Et	CH_2OEt	S	8.9276	0.1660
39	4-Ph	Et	CH_2Ph	O	8.7020	0.1674
40	4-Ph	Et	CH_2Ph	S	9.0633	0.2103
41	4-Ph	Et	Ph	O	9.6316	0.2551
42	4-Ph	Et	Ph	S	10.4771	0.2218
43	2-NH_2	i-Pr	CH_2Ph	O	8.1110	0.2869
44	2-NH_2	i-Pr	CH_2Ph	S	8.4838	0.1413
45	3-NH_2	i-Pr	CH_2Ph	O	8.5363	0.3815
46	3-NH_2	i-Pr	CH_2Ph	S	8.9027	0.2169
47	4-NH_2	i-Pr	CH_2Ph	O	9.0540	0.4448
48	4-NH_2	i-Pr	CH_2Ph	S	9.4257	0.2513
49	2-OH	i-Pr	CH_2Ph	O	8.1989	0.1886
50	2-OH	i-Pr	CH_2Ph	S	8.5728	0.0987
51	3-OH	i-Pr	CH_2Ph	O	8.5031	0.2365
52	3-OH	i-Pr	CH_2Ph	S	8.8729	0.1314
53	4-OH	i-Pr	CH_2Ph	O	7.9001	0.1590
54	4-OH	i-Pr	CH_2Ph	S	8.2718	0.0929
55	2-F	i-Pr	CH_2Ph	O	8.5049	0.1162
56	2-F	i-Pr	CH_2Ph	S	8.8718	0.1077
57	3-F	i-Pr	CH_2Ph	O	8.7381	0.1419
58	3-F	i-Pr	CH_2Ph	S	9.1120	0.1213
59	4-F	i-Pr	CH_2Ph	O	8.1364	0.0930
60	4-F	i-Pr	CH_2Ph	S	8.5090	0.1108

(*Continued to next page*)

Table 6. Continued

A/A	R_1	R_2	R_3	X	log (1/IC_{50}) pred	Leverages (limit 0.30)
61	2-Cl	i-Pr	CH_2Ph	O	8.9507	0.1524
62	2-Cl	i-Pr	CH_2Ph	S	9.3248	0.1939
63	3-Cl	i-Pr	CH_2Ph	O	8.9205	0.1181
64	3-Cl	i-Pr	CH_2Ph	S	9.2949	0.1521
65	4-Cl	i-Pr	CH_2Ph	O	8.3577	0.0945
66	4-Cl	i-Pr	CH_2Ph	S	8.7308	0.1662
67	2-Br	i-Pr	CH_2Ph	O	8.8951	0.1267
68	2-Br	i-Pr	CH_2Ph	S	9.2652	0.1839
69	3-Br	i-Pr	CH_2Ph	O	8.9465	0.1129
70	3-Br	i-Pr	CH_2Ph	S	9.3209	0.1616
71	4-Br	i-Pr	CH_2Ph	O	8.4096	0.0982
72	4-Br	i-Pr	CH_2Ph	S	8.7832	0.1842
73	3,5-$(NH_2)_2$	i-Pr	CH_2Ph	O	8.9609	0.8533
74	3,5-$(NH_2)_2$	i-Pr	CH_2Ph	S	9.3245	0.5330
75	3,5-$(OH)_2$	i-Pr	CH_2Ph	O	9.0008	0.4640
76	3,5-$(OH)_2$	i-Pr	CH_2Ph	S	9.3676	0.2598
77	3,5-$(Br)_2$	i-Pr	CH_2Ph	O	9.8257	0.1562
78	3,5-$(Br)_2$	i-Pr	CH_2Ph	S	10.2029	0.2571
79	3,5-$(Cl)_2$	i-Pr	CH_2Ph	O	9.7759	0.1523
80	3,5-$(Cl)_2$	i-Pr	CH_2Ph	S	10.1526	0.2236
81	3,5-$(F)_2$	i-Pr	CH_2Ph	O	9.4312	0.1927
82	3,5-$(F)_2$	i-Pr	CH_2Ph	S	9.8067	0.1545

and test) shows that our MLR model can predict external data with great accuracy. The proposed method, due to the high predictive ability, is a useful aid to the costly and time consuming experiments for determining anti-HIV activity. The proposed models can be used to screen existing databases or virtual libraries in order to identify novel potent compounds. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" compounds. An attempt in this direction was carried out. Synthesis of the new proposed molecules and their biological evaluation will show if this procedure can be used as a general rational drug discovery tool.

Acknowledgements

The authors wish to thank Professor Alexander Tropsha for his valuable suggestions that helped to improve the significance and rigor of the manuscript. A.A. wishes to thank Cyprus Research Promotion Foundation (Grant No. PENEK/ENISX/0603/05) and A.G. Leventis Foundation for its financial support. G.M. thanks the Empirikion Foundation for financial support.

References

1. Hansh, C. and Leo, A. Exploring QSAR. *Fundamentals and Applications in Chemistry and Biology.* American Chemical Society. Washington, DC, 1995.
2. Miyasaka, T., Tanaka, H., Baba, M., Hayakawa, H., Walker, R., Balzarini, J. and Clercq, E. *A novel lead for specific anti-HIV-1 agents: 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine.* J. Med. Chem., 32 (1989) 2507–2509.
3. Hansch, C. and Zhang, L. Bioorg. *QSAR of HIV inhibitors.* Med. Chem. Lett., 2 (1992) 1165–1169.
4. Hannongbua, S., Lawtrakul, L. and Limtrakul, *J. Structure – Activity Correlation Study of HIV-1 Inhibitors. Electron and Molecular Parameters.* J. Comput. – Aided Mol. Des., 10 (1996) 145–152.
5. Tanaka, H., Takashima, H., Ubasawa, M., Sekiya, K., Nitta, I., Baba, M., Shigeta, Sh., Walker, R.T. De Clercq, E., Miyasaka, T. *Synthesis and Antiviral Activity of 6-Benzyl Analogs of 1-[(2-Hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents.* J. Med. Chem., 35 (1992) 4713–4719.
6. Luco, J.M. and Ferreti F.H. *QSAR Based on Multiple Linear Regression and PLS Methods for the Anti-HIV Activity of a Large Group of HEPT Derivatives.* J. Chem. Inf. Comput. Sci., 37 (1997) 392–401.
7. Jalali-Heravi., M. and Parastar., F. J. *Use of Artificial Neural Networks in a QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives*, Chem. Inf. Sci., 40 (2000), 147–154.
8. Alves, C. N., Pinheiro, J. C. Camargo, A. J., Ferreira, M. M. C. and Silva, A. B. F. *A structure – activity relationship study of HEPT – analog compound with anti – HIV activity.* Journal of Molecular Structure (Theochem), 530 (2000) 39–47.
9. Bazoui, H., Zahouily, M., Boulajaaj, S., Sebti, S. and Zakarya, D. *QSAR for anti-HIV activity of HEPT derivatives.* SAR and QSAR in Environmental Research, 13 (2002) 567–577.
10. Duali, L., Villemin, D. and Cherqaoui, D., *Neural Networks : Accurate Nonlinear QSAR Model for HEPT Derivatives*, J. Chem. Inf. Comput. Sci., 43 (2003) 1200–1207.
11. Duali, L., Villemin, D. and Cherqaoui, D., *Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives.* Curr. Pharm. Des., 9 (2003) 1817–1826.
12. Duali, L., Villemin, D., Zyad, A. and Cherqaoui, D., *Artificial neural networks: non-linear QSAR studies of HEPT derivatives as HIV-1 reserve transcriptase inhibitors.* Molecular Diversity, 8 (2004) 1–8.
13. Gupta, S., Singh, M. and Madam, A. K., *Predicting anti-HIV activity: computational approach using a novel topological descriptor.* J. Comput. Aided Mol. Des., 15 (2001) 671–678.

14. Gayen, S., Debnath, B., Samanta, S. and Jha, T., *QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters.* Bioorganic & Medicinal Chemistry, 12 (2004) 1493–1503.
15. Efroymson, M. A., *Multiple Regression Analysis*, in: Ralston, A. and Wilf, H.S. (eds.), Mathematical Methods for Digital Computers, Wiley, New York, NY, 1960.
16. Osten, D. W. *Selection of Oprimal Regression Models via Cross-Validation* J. Chemom., 2 (1988) 39–48
17. Wold, S. and Eriksson, L., *Statistical Validation of QSAR Results*, in: Van de Waterbeemd, H. (Ed.), Chemometrics Methods in Molecular Design, VCH Weinheim (Germany), 1995, pp. 309–318.
18. Tropsha, A., Gramatica, P. and Gombar, V. K., *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*. Quantitative Structure Activity Relationships, 22 (2003) 1–9.
19. Golbraikh, A. and Tropsha, A. *Beware of q^2!* Journal of Molecular Graphics and Modelling, 20 (2002) 269–276.
20. Shen, M., Beguin, C., Golbraikh, A., Stables, J., Kohn, H. and Tropsha, A. *Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds.* J. Med. Chem., 47 (2004) 2356–2364.
21. Atkinson, A. C., *Plots, transformations and regression*, Clarendon Press, Oxford (UK), 1985, p. 282.
22. Gulyaeva, N., Zaslavsky, A., Lechner, P., Chlenov, M., Chait, A., Zaslavsky, B. *Relative hydrophobicity and lipophilicity of β-blockers and related compounds as measured by aqueous two-phase partitioning, octanol-buffer partitioning, and HPLC.* Eur. J. Pharm. Sci., 17 (2002) 81–93.
23. Walters, W. P., Ajay, Murcko, M. A. *Recognizing molecules with drug-like properties.* Curr. Opin. Chem. Biol. 3 (1999) 384–387.
24. Devillers, L. (Ed.), Comparative QSAR. Taylor and Francis, Washington, DC. 1998.
25. Golbraikh, A. and Tropsha, A., *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*. Molecular Diversity, 5 (2000) 231–243.
26. Vanyur, R., Heberger, K., Jakus, J., *Prediction of Anti-HIV-1 Activity of a Series of Tetrapyrrole Molecules*. Journal Chemical Information Computer Science, 43 (2003) 1829–1836.