Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS

J. Serb. Chem. Soc. 78 (4) 495–506 (2013)
JSCS–4433

UDC 616.98:578.828:547.568–36+
547.854.4:615.37
Original scientific paper

A quantitative structure–activity relationships study for the anti-HIV-1 activities of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine derivatives using the multiple linear regression and partial least squares methodologies

DANIELA IVAN, LUMINITA CRISAN*, SIMONA FUNAR-TIMOFEI
and MIRCEA MRACEC

*Institute of Chemistry of Romanian Academy, Department of Computational Chemistry,
24 Mihai Viteazul Bvd., 300223, Timisoara, Romania*

(Received 13 July 2012)

Abstract: A quantitative structure–activity relationships (QSAR) study using Multiple Linear Regression (MLR) and Partial Least Squares (PLS) methodologies was performed for a series of 127 derivatives of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT), a potent inhibitor of the of the human immunodeficiency virus type 1, HIV-1 reverse transcriptase (RT). The MLR and PLS methods were employed to explore the relationship between the descriptors (as independent variables) of a pool of HEPT derivative and anti-HIV-1 activity, expressed as log (1/*EC*$_{50}$) (as dependent variables). Using Dragon descriptors, the present study was aimed at developing a predictive and robust QSAR model for predicting anti-HIV activity of HEPT derivatives for a better understanding of the molecular features of these compounds important for their biological activity. According to the squared correlation coefficients, which had values between 0.826 and 0.809 for the MLR and PLS methods, the results demonstrated almost identical qualities and good predictive ability for both the MLR and PLS models. After dividing the dataset into training and test sets, the model predictability was tested by several parameters, including the Golbraikh–Tropsha external criteria and the goodness of fit, tested using the *Y*-randomization test.

Keywords: Golbraikh–Tropsha criteria; Dragon descriptors; *Y*-randomization.

INTRODUCTION

Infection with the human immunodeficiency virus type-1 (HIV-1) causes increasing destruction of immunity, which finally results in the development of the immunodeficiency syndrome (AIDS). HIV-1 reverse transcriptase (RT) is one of the enzymes responsible for the replication of HIV-1.[1] The majority of com-

* Corresponding author. E-mail: lumi_crisan@acad-icht.tm.edu.ro

doi: 10.2298/JSC120713085I

pounds used in the treatment of the HIV-1 infections are inhibitors of the reverse transcriptase. They belong to two main classes: 1) analogues of 2′,3′-dideoxynucleoside (ddNs), such as: zidovudine, didanosine, lamivudine and 2) non-nucleoside RT inhibitors (NNRTs), such as nevirapine, delavirdine, tivirapine, MKC-442 and 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives.[2–4] The HEPT derivatives are among the most selective non-nucleoside drugs discovered and were tested on MT-4 cells by Tanaka and coworkers.[5,6] Due to their high specificity and low toxicity, these inhibitors are promising candidates for the treatment of AIDS, having been extensively studied for many years.[7]

Quantitative structure–activity relationships (QSAR) studies were previously performed on HEPT derivatives used as HIV-1 reverse transcriptase inhibitors by several methods, *i.e.*, Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) applied to 90[8] and 103[9] HEPT derivatives, respectively, Minimum Steric Difference (MTD) to 34 compounds,[10] Neural Networks (NN) to 80[11] and 103[12] compounds, Particle Swarm Optimization (PSO) and Support Vector Machine (SVM) to 40 compounds,[13] Partial Least Squares (PLS) to 107 compounds,[14] Genetic Programming (GP) to 80 compounds,[15] MLR to 79[16] and 71[17] HEPT derivatives and Comparative Molecular Field Analysis (CMFA) to 44 derivatives[18]. In QSAR, molecular descriptors are correlated with the biological activity of different series of compounds, with the purpose of investigating their binding mechanism. QSAR analysis can indicate which features of a given molecule enable the design of new and more potent compounds with strengthened biological activities.[19] The MLR and the PLS approaches are the most used computational methods in QSAR studies. These techniques can be employed to better interpret the pharmacological data and to predict new biologically active compounds.[20]

An analysis using the MLR and PLS methods applied to a series of 127[12,14] HEPT derivatives with known biological activity has not yet been published in the literature. In this study, these approaches were applied to these compounds and it was found that the anti-HIV-1 activities could be significantly described by Dragon descriptors.[21] The purpose of this paper is to offer a contribution to the understanding of the influence of the molecular features of these 127 HEPT derivatives on their anti-HIV-1 activity using QSAR procedures.

MATERIAL AND METHODS

Compounds studied

A set of 127 HEPT compounds (having the common skeleton presented in Fig. 1) with known biological activity (presented in Table S-I, Supplementary material to this paper) was analyzed in this study. The anti-HIV activity data (A_{obs}), expressed as log ($1/EC_{50}$), where EC_{50} represents the concentration that produces a 50 % protection of MT-4 cells against the cytopathic effect of HIV-1, were taken from the literature.[12,14]

Structural calculations

In the first step, the structures of the 127 investigated molecules were pre-optimized using the (MM+) Molecular Mechanics Force Field included in the HyperChem 7.52 package.[22] In the next step, the minimized structures were refined using the semi-empirical AM1 Hamiltonian also implemented in HyperChem. For geometry optimization, a gradient norm limit of 0.01 kcal Å^{-1} was set. To display the "real" spatial orientation of the substituents of the HEPT derivatives, all compounds investigated were superposed on the X-ray coordinates of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (http://www.rcsb.org/pdb/explore/ /explore.do?structureId=1RTI). The RMS fit criterion was calculated for the superposition of three atoms included in the rigid skeleton, namely N3, C6, C1' (Fig. 1, atom numbering as per the IUPAC convention).

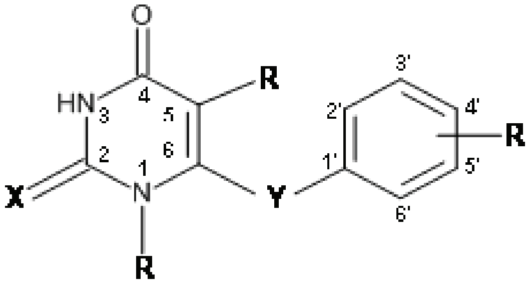

Fig. 1. General structure of the HEPT derivatives.

Molecular descriptors

1217 molecular descriptors (0D, 1D, 2D, 3D) were used to characterize the set of 127 potent inhibitors of the HIV-1 reverse transcriptase. The descriptors calculated using the DRAGON[23] software were analyzed to check and remove constant or near-constant variables. The remaining descriptors were used to build the *X*-matrix in the MLR and PLS analysis, as follows:[21] 35 constitutional descriptors, 92 topological descriptors, 42 walk and path counts, 31 connectivity indices, 47 information indices, 92 2D autocorrelations, 63 Burden eigenvalue descriptors, 18 topological charge indices, 40 eigenvalue-based indices, 41 Randic molecular profiles, 53 geometrical descriptors, 133 RDF descriptors, 160 3D-MoRSE descriptors, 99 Weighted Holistic Invariant Molecular (WHIM) descriptors, 195 Geometry, Topology and Atom-Weights Assembly (GETAWAY) descriptors, 22 functional group counts, 35 atom-centered fragments, 14 charge descriptors, and 5 molecular properties.

Training and test set generation

The clustering technique is intensively applied to split data sets into training and test sets to perform further QSAR modeling.[24] Therefore the calculated Dragon descriptors were normalized and introduced in the R package[25] for clustering. The HEPT derivatives were divided into training and test sets by means of the partition against medoids (PAM) algorithm.[26] The diversity criteria used to estimate the dissimilarity of molecules was the Euclidean distance. The initial dataset of 127 HEPT derivatives was split into fifteen clusters following the maximum silhouette value criteria. The training/test sets were built taking randomly 25 % of each cluster as the test set, while the remaining 75 % were used as the training set. In the situation of singletons, the compounds have been designated for the training set, but the equilibration of the test set was performed by the inclusion of supplementary compounds from larger clusters.

MLR method

As the number of calculated descriptors of 1217 is too high compared to the number of compounds (N = 127), an appropriate variable selection method was necessary. The Genetic Algorithm (GA) is a reliable and extensively used variable selection method.[27,28] GA uses a stochastic algorithm that elucidates the optimization issues illustrated by fitness criteria, involving the evolution assumption of Darwin and various genetic functions, including crossover and mutation. The MobyDigs[29] package uses GAs to select the significant descriptors that influence the variation of biological activity of the compounds studied in this work. In MobyDigs, the following parameters were used: the RQK fitness function[30] with leave-one-out cross-validation[31] correlation coefficient as constrained function to be optimized, a crossover/mutation trade-off parameter of T = 0.5 and a model population size of P = 50.

PLS method

The PLS method is a statistical modeling technique that simultaneously works with two matrices, **X** (dependent variables, *e.g.*, molecular descriptors) and **Y** (independent variables, *e.g.*, biological activity) to model the relationship between them.[32] The relationship between the **X** and **Y** matrices is described as a latent variable approach and is preferable for large data sets.[33] The main advantage of this method in comparison to MLR is that interrelated variables can be included in the model. This could lead to a stable and highly predictive model.[34] Using the SIMCA-P+ 12.0 package,[35] the QSAR matrix (including the **X** and **Y** matrices) was analyzed in the first step by Principal Component Analysis (PCA),[36] and subsequently by Partial Least Squares (PLS)[37] approaches. The PLS method is especially useful for larger data sets. The squared correlation regression coefficient R^2, and the squared cross-validated correlation coefficient, Q^2, are the most important statistical parameters that provide a measure of the quality and validity for the final PLS model, while the Variables Importance in the Projection (VIP) values and the sign of the coefficients of the variables are more relevant in explaining the activity mechanism. The significant principal components were selected by 7 cross-validation groups.

Model validity

The predictability of the model was tested with the Golbraikh–Tropsha[38-40] criteria and the goodness of fit with the Y-randomization test.[41] The following Golbraikh–Tropsha conditions should be satisfied to certify the predictive ability of the MLR and PLS models:

i) $Q^2 > 0.5$ (squared cross-validation correlation coefficient);

ii) $R^2 > 0.6$ (the squared correlation coefficient R between the predicted and observed activities);

iii) $(R^2 - R_0^2) / R^2 < 0.1$ or $(R^2 - R_0'^2) / R^2 <0.1$ and $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$ (concerning the coefficients of determination for the predicted *vs*. the observed activities R_0^2, and the observed *vs*. predicted activities $R_0'^2$ through the origin and slopes k and k' of the regression lines through the origin);

iv) $\left|R_0^2 - R_0'^2\right| < 0.3$.

The predictive power of QSAR models is frequently judged based on the predictive parameter R^2 (R^2_{pred}).[42] For a predictive QSAR model, the value of R^2_{pred} (presented in Eq. (1)) should be higher than 0.5:[24,42]

$$R^2_{\text{pred}} = 1 - \frac{\sum\left(Y_{\text{pred(test)}} - Y_{\text{(test)}}\right)^2}{\sum\left(Y_{\text{(test)}} - \bar{Y}_{\text{training}}\right)^2} \tag{1}$$

The *Y*-randomization test is a widely used technique that displays the robustness of a QSAR model, being a measure of the model overfit. The dependent variable (biological activity) is randomly shuffled and a QSAR model is built using the same descriptor matrix. The obtained MLR and PLS models (after 500 randomizations) must have the minimal R^2 and Q^2 values.[42]

RESULTS AND DISSCUSION

MLR analysis

Using the above-mentioned genetic algorithm, the best MLR Equation (2) was obtained:

$$\begin{gathered} A_i = 33.915(\pm 5.594) - 0.155(\pm 0.011)DELS - 43.507(\pm 8.332)X0A + \\ 3.724(\pm 0.254)GGI3 + 5.056(\pm 1.072)GATS1v - 3.075(\pm 0.649)R4u \end{gathered} \tag{2}$$

$$N_{\text{training}} = 95, N_{\text{test}} = 32; R^2 = 0.826; Q^2_{\text{LOO}} = 0.802; Q^2_{\text{boot}} = 0.788;$$

$$Q^2_{\text{ext}} = 0.618; a(R^2) = 0.156; a(Q^2) = 0.039; R^2_{\text{adj}} = 0.816; SDEC = 0.66;$$

$$F = 84.4; s = 0.682; AIC = 0.528; Kx = 32.21; Kxy = 38.06$$

where R^2 represents the correlation coefficient, Q^2_{LOO} – leave-one-out cross-validation parameter, Q^2_{boot} – bootstrapping parameter,[31] Q^2_{ext} – external Q^2,[31] $a(R^2)$ and $a(Q^2)$ – *Y*-scrambling variables,[31] R^2_{adj} – adjusted R^2, *SDEP* – standard deviation error in the prediction, *SDEC* – standard deviation error in the calculation,[31] *F* – Fischer test, *s* – standard error of estimate, *AIC* – Akaike information criterion,[31] the multivariate *K* – correlation indices (*Kx* – the multivariate correlation index of the matrix of *X* descriptors and *Kxy* – the multivariate correlation index of the matrix of *X* descriptors and *Y* response variables).[31]

In the present study, the best MLR model had five parameters. A higher or lower number of molecular descriptors did not have any significant effect on the accuracy of the model. Additionally, the predictive R^2 (leave-one-out cross validation parameter, Q^2_{LOO}) and external Q^2 (Q^2_{ext}) values were calculated and are presented in Table I. The most important descriptors (Table S-II in the Supple-

TABLE I. Correlation matrix of the five selected descriptors included in E 2; *DELS* represents the molecular electrotopological variation; *X0A* represents the average connectivity index of order 0; *GGI*3 represents the topological charge index of order 3; *GATS1v* represents the Geary autocorrelation of lag 1 weighted by the van der Waals volume; *R4u* represents the *R* autocorrelation of lag 4 / unweighted

	DELS	*X0A*	*GGI*3	*GATS*1v	*R*4*u*
DELS	1.000				
X0A	0.182	1.000			
*GGI*3	0.056	0.477	1.000		
*GATS*1v	–0.099	0.295	0.119	1.000	
*R*4*u*	–0.215	0.261	0.372	0.512	1.000

mentary material to this paper), selected by a genetic algorithm, which influence the anti-HIV activity are the topological *DELS* descriptor (which describes the molecular electrotopological variation), the *X0A* (average connectivity index chi-0) connectivity index, the *GGI*3 (topological charge index of order 3 topological charge index, the *GATS*1*v* (Geary autocorrelation of lag 1/weighted by the van der Waals volume) 2D autocorrelations and the *R*4*u* (*R* autocorrelation of lag 4/unweighted) GETAWAY descriptors.[21]

An intercorrelation analysis of the selected molecular descriptors performed with the Statistica software[43] is presented in Eq. (2) (Table I). The five selected descriptors are not intercorrelated.

The statistical results and intercorrelation coefficients presented in Eq. (2) and Table I, which the MLR method associated with a proper variable selection procedure, generates an efficient QSAR model for predicting the anti-HIV-1 activity of the different HEPT derivatives.

PLS analysis

A PCA model was built with the SIMCA-P+ version 12.0[35] software for the whole **X** matrix (including $N = 127$ compounds and $X = 1217$). From the 63 significant principal components resulting from this analysis, the first three components already explained 58.2 % of the information content of the descriptor matrix. PLS calculations were also performed by the same program using 95 HEPT derivatives as a training set and 32 compounds as a test set. The statistical results of the PLS model: $R^2{}_Y$(CUM) = 0.864 and Q^2(CUM) = 0.794 obtained for the four principal components demonstrated the model overfit. This inconvenience was overcome by excluding the noise variables from this model (*i.e.*, the variables with coefficient values insignificantly different from 0). Thus, a robust model, M17 ($N = 95$ and $X = 63$) with one latent variable (Table II) was obtained. Thus, the predictive power of the final PLS model was tested in the next step using the Golbraikh–Tropsha criteria, and the R^2_{pred} tests.

TABLE II. Statistical characteristics of the final PLS model; $R^2_{X\,\text{(CUM)}}$ and $R^2_{Y\text{(CUM)}}$ are the cumulative sum of squares of all the *X* and *Y* values, respectively, explained by all extracted principal components; $Q^2_{\text{(CUM)}}$ is the fraction of the total variation of the *Y* values that can be predicted for all the *A* extracted principal components in the cross-validation procedure (7 rounds) used to establish the number of significant principal components, *A*

PLS model	$R^2_{X\,\text{(CUM)}}$	$R^2_{Y\text{(CUM)}}$	Q^2(CUM)	*N*	*A*	*X*
M17	0.471	0.809	0.803	95	1	63

The descriptors summarized by the significant first principal component in M17 explain 47.1 % of the variation. 31 of the 63 selected variables in M17 (Table III) had *VIP* values greater than 1 and were considered to be the most relevant for the model.

TABLE III. The coefficients in descending order of *VIP* values for the first principal component of the M17 model

No.	Variable ID	*CoefCS* [1]	*VIP* [1]	Descriptor significance
1	MATS5e[a]	–0.0271	1.303	Moran autocorrelation of lag 5 weighted by the Sanderson electronegativity
2	PW4[b]	0.0262	1.257	Path/walk 4 – Randic shape index
3	GATS5e[a]	0.0256	1.231	Geary autocorrelation of lag 5 weighted by the Sanderson electronegativity
4	X3A[c]	–0.0248	1.190	Average connectivity index of order 3
5	Ms[d]	–0.0247	1.186	Mean electrotopological state
6	Me[c]	–0.0242	1.162	Mean atomic Sanderson electronegativity (scaled on the carbon atom)
7	CIC3[e]	0.0241	1.157	Complementary Information Content index (neighborhood symmetry of 3rd order)
8	SPAM[f]	–0.0238	1.143	average span *R*
9	SIC3[e]	–0.0238	1.141	Structural Information Content index (neighborhood symmetry of 3rd order)
10	ATS5p[a]	0.0237	1.139	Broto–Moreau autocorrelation of lag 5 (log function) weighted by polarizability
11	BIC4[e]	–0.0236	1.131	Bond Information Content index (neighborhood symmetry of 4th order)
12	CIC4[e]	0.0235	1.129	Complementary Information Content index (neighborhood symmetry of 4th order)
13	AlogP[g]	0.0234	1.123	Ghose–Crippen octanol–water partition coeff. (log *P*)
14	MATS7e[a]	–0.0233	1.117	Moran autocorrelation of lag 7 weighted by the Sanderson electronegativity
15	MATS5m[a]	–0.0233	1.117	Moran autocorrelation of lag 5 weighted by mass
16	GATS5m[a]	0.0232	1.115	Geary autocorrelation of lag 5 weighted by mass
17	MATS6e[a]	0.0232	1.113	Moran autocorrelation of lag 6 weighted by the Sanderson electronegativity
18	nHAcc[h]	–0.0232	1.113	Number of acceptor atoms for H-bonds (N,O,F)
19	SIC4[e]	–0.0231	1.108	Structural Information Content index (neighborhood symmetry of 4th order)
20	GATS7e[a]	0.0230	1.105	Geary autocorrelation of lag 7 weighted by the Sanderson electronegativity
21	SEigv[i]	0.0229	1.101	Eigenvalue sum from the van der Waals weighted distance matrix
22	R8p[j]	0.0227	1.090	*R* autocorrelation of lag 8 / weighted by polarizability
23	BIC5[e]	–0.0227	1.088	Bond Information Content index (neighborhood symmetry of 5th order)
24	SEigp[i]	0.0225	1.081	Eigenvalue sum from polarizability weighted distance matrix

TABLE III. Continued

No.	Variable ID	*CoefCS* [1]	*VIP* [1]	Descriptor significance
25	ATS5v[a]	0.0225	1.079	Broto–Moreau autocorrelation of lag 5 (log function) weighted by the van der Waals volume
26	SEige[i]	–0.0223	1.072	Eigenvalue sum from electronegativity weighted by the distance matrix
27	Mor04m[k]	–0.0223	1.071	signal 04 / weighted by mass
28	ATS4v[a]	0.0221	1.061	Broto–Moreau autocorrelation of lag 4 (log function) weighted by the van der Waals volume
29	Mor04p[k]	–0.0220	1.056	signal 04 / weighted by the polarizability
30	nO[b]	–0.022	1.055	Number of oxygen atoms
31	Hy[g]	–0.021	1.040	Hydrophilic factor

[a]2D autocorrelations; [b]topological descriptors; [c]connectivity indices; [d]constitutional descriptors; [e]information indices; [f]geometrical descriptors; [g]molecular properties; [h]functional group counts; [i]eigenvalue-based ondices; [j]GETAWAY descriptors; [k]3D-Morse descriptors

Model validation results

The predictive abilities of the best MLR and PLS models were tested (Table IV) using the Golbraikh–Tropsha criteria and the R^2_{pred} test (see Model Validity section). All the calculated parameters indicated that both models showed a good predictive power.

TABLE IV. Predictive power results for the external test set; Golbraikh and Tropsha criteria were used

Model	M1 (MLR)	M17 (PLS)
R^2	0.675	0.717
R^2_0	0.655	0.714
R'^2_0	0.635	0.952
k	1.000	0.937
k'	0.974	1.000
$\lvert R^2_0 - R'^2_0 \rvert$	0.02	0.238
$\frac{R^2 - R^2_0}{R^2}$	0.029	0.042
$\frac{R^2 - R'^2_0}{R^2}$	0.059	–0.61
R^2_{pred}	0.663	0.715

Analyzing the results of the external test set listed in Table IV, it could be observed that all the Golbraikh–Tropsha criteria were fulfilled.

The dependence between the observed (A_{obs}) *vs*. predicted (A_{pred}) anti-HIV activity values is presented in Fig. 2a for the final MLR model and in Fig. 2b for the final PLS model.

The *Y*-randomization procedure was applied to the test set by SIMCA-P+ +12.0 software[35] (for the final PLS model) and by the MobyDigs program[29] (for the final MLR model). This gave the following intercept (PLS/MLR) values of the regression lines obtained by the correlation between the calculated R^2, respectively Q^2 values of the original *Y*-variable and the shuffled *Y*-variable, respectively: –0.00267/0.159 for the R_Y^2 line and –0.125/0.039 for the Q_Y^2 line. The slope values close to zero indicate stable models.

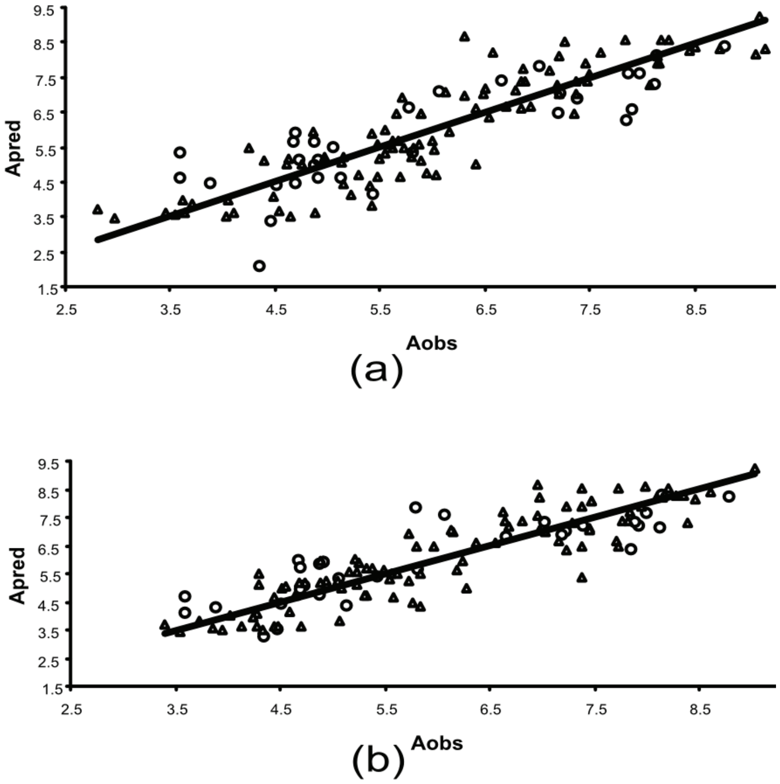

Fig. 2. Plots of the predicted (A_{pred}) *vs*. observed (A_{obs}) anti-HIV activity for the training set (triangles) and the test set (circles) for the final MLR (a) and PLS (b) model.

CONCLUSIONS

The final models obtained using the MLR and PLS methods had good statistical parameters and excellent predictive capacity.

The most important molecular descriptors included in the final MLR and PLS models are related to the geometric representation of the molecules, providing information on the interatomic distances, topological distances, types of atoms (in case of 2D-autocorrelation descriptors), information derived from the molecular graph in 2D space, *i.e.*, the connectivity, counting paths, walks, vertex degree of the atoms, *etc*. (expressed by connectivity indices and topological descriptors) and 3D information (by the GETAWAY parameters).

SUPPLEMENTARY MATERIAL

The chemical structure of the studied HEPT derivatives, predicted anti-HIV-1 activity values and the selected molecular descriptors included in the final MLR model, and predicted activity value obtained by the PLS method, are available electronically from http:// //www.shd.org.rs/JSCS/, or from the corresponding author on request.

Acknowledgements. This project was financially supported by Project 1.1 and 1.2 of the Institute of Chemistry of the Romanian Academy. Statistica, MobyDigs and Simca-P+ acquisition was funded by Ministerul Educatiei, Cercetarii si Tineretului – Autoritatea Nationala pentru Cercetare Stiintifica (MedC-ANCS), Romania, Contract Grant No.: 71GR/2006.

ИЗВОД

QSAR МОДЕЛОВАЊЕ АНТИ-HIV-1 АКТИВНОСТИ ДЕРИВАТА 1-[(2-ХИДРОКСИЕТОКСИ)МЕТИЛ]-6-(ФЕНИЛТИО)ТИМИНА ПРИМЕНОМ МЕТОДОЛОГИЈА ВИШЕСТРУКЕ ЛИНЕАРНЕ РЕГРЕСИЈЕ И ПАРЦИЈАЛНИХ НАЈМАЊИХ КВАДРАТА

DANIELA IVAN, LUMINITA CRISAN, SIMONA FUNAR-TIMOFEI и MIRCEA MRACEC

Institute of Chemistry of Romanian Academy, Timisoara, Romania

Изведено је QSAR моделовање применом методологије вишеструке линеарне регресије (MLR) и парцијалних најмањих квадрата (PLS) на серији од 127 деривата 1-[(2--хидроксиетокси)метил]-6-(фенилтио)тимина (HEPT), који су моћни инхибитори реверсне транскриптазе (RT) вируса хумане иммунодефициенције типе HIV-1. Да би се истражила веза између дескриптора структуре HEPT деривата (као независних променљивих) и анти-HIV-1 активности, изражене преко $\log(1/EC_{50})$ вредности, примењене су MLR и PLS методе. На основу квадрата коефицијената корелације, чије вредности леже између 0,826 и 0,809, закључено је да обе методе (MLR и PLS) имају скоро идентичну моћ предвиђања.

(Примљено 13. јула 2012)

REFERENCES

1. M. E. Goldman, J. H. Nunberg, J. A. O'Brien, J. C. Quintero, W. A. Schleif, K. F. Freund, S. Lee Gaul, W. S. Saari, J. S. Wai, J. M. Hoffman, P. S. Anderson, D. J. Hupe, E. A. Emin, *Proc. Natl. Acad. Sci. U.S.A.* **88** (1991) 6863
2. E. De Clercq, *Pure Appl. Chem.* **73** (2001) 55
3. M. Baba, H. Tanaka, E. De Clercq, R. Pauwels, J. Balzarini, D. Schols, H. Nakashima, C. F. Perno, R. T. Walker, T. Miyasaka, *Biochem. Biophys. Res. Commun.* **165 (**1989) 1375
4. T. Miyasaka, H. Tanaka, R. T. Walker, J. Balzarini, E. De Clercq, *J. Med. Chem.* **32** (1989) 2507
5. H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, R. T. Walker, E. DeClercq, T. Miyasaka, *J. Med. Chem.* **35** (1992) 337
6. H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, R. T. Walker, E. DeClercq, T. Miyasaka, *J. Med. Chem.* **35** (1992) 4713

7. H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, N. Inouye, M. Baba, S. Shigeta, R. T. Walker, E. DeClercq, T. Miyasaka, *J. Med. Chem.* **38** (1995) 2860
8. H. Bazoui, M. Zahouily, S. Sebti, S. Boulajaaj, D. Zakarya, *J. Mol. Modell.* **8** (2002) 1
9. H. Bazoui, M. Zahouily, S. Boulajaaj, S. Sebti, D. Zakarya, *SAR QSAR Environ. Res.* **13** (2002) 567
10. M. Mracec, D. Ivan, M. Mracec, *Rev. Roum. Chim*. **49** (2004) 431
11. L. Douali, D. Villemin, A. Zyad, D. Cherqaoui, *Mol. Diversity* **8** (2004) 1
12. M. Zahouily, J. Rakik, M. Lazar, M. A. Bahlaoui, A. Rayadh, N. Komiha, *ARKIVOC* (2007) 245
13. L. Lawtrakul, C. Prakasvudhisarn, *Monatsh. Chem*. **136** (2005) 1681
14. A. Bak, J. Polanski, *Bioorg. Med. Chem.* **14** (2006) 273
15. M. Arakawa , K. Hasegawa , K. Funatsu, *Chemom. Intell. Lab. Syst*. **83** (2006) 91
16. C. Duda-Seiman, D. Duda-Seimana, M. V. Putz, D. Ciubotariu, *Dig. J. Nanomater. Bios.* **2** (2007) 207
17. V. Ravichandran, V. K. Mourya, R. K. Agrawal, *Dig. J. Nanomater. Bios.* **3** (2008) 9
18. R. S. Latha, R. Vijayaraj, E. R. Singam, K. Chitra, V. Subramanian, *Chem. Biol. Drug. Des.* **78** (2011) 418
19. A. Afantitis, G. Melagraki, H. Sarimveis, O. Igglessi-Markopoulou, J. Markopoulos, P.A. Koutentis, *Mol. Diversity* **10** (2006) 405
20. A. Ganjee , X. Lin, *J. Med. Chem*. **48** (2005) 1448
21. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, New York, USA, 2009
22. HYPERCHEM 7.52, release for windows; hypercube, Inc., Gainesville, FL, USA, http://www.hyper.com
23. DRAGON, version 3.0, 2003 for windows (software for molecular descriptor calculations), http://www.talete.mi.it
24. J. T. Leonard, K. Roy, *SAR Comb. Sci.* **25** (2006) 235
25. R Development Core Team 2010, ISBN 3-900051-07-0, available at www.r-project.org
26. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990
27. U. Depczynski, V. J. Frost, K. Molt, *Anal. Chim. Acta* **420** (2000) 217
28. B. K. Alsberg, N. Marchand-Geneste, R. D. King, *Chemom. Intell. Lab. Syst.* **54** (2000) 75
29. R. Todeschini, V. Consonni, A. Mauri, M. Pavan, in *Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*, R. Leardi, Ed., Elsevier, Amsterdam, 2004, p.141
30. R. Todeschini, V. Consonni, A. Mauri, M. Pavan, *Anal. Chim. Acta* **515** (2004) 199
31. D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **43** (2003) 579
32. H. Wold, in *Multivariate analysis*, P. R. Krishnaiah, Ed., Academic Press, New York, 1966, p. 391
33. L. Eriksson, J. Gottfries, E. Johansson, S. Wold, *Chemom. Intell. Lab. Syst.* **73** (2004) 73
34. A. Höskuldsson, *J. Chemom.* **2** (1988) 211
35. SIMCA-P+, version 12.0, Umetrics AB: Umea, Sweden, http://www.umetrics.com
36. M. Daszykowski, K. Kaczmarek, V. Heyden, B. Walczak, *Chemom. Intell. Lab. Syst.* **85** (2007) 203
37. H. Wold, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz, N. L. Johnson, Eds., Wiley, New York, 1985, p. 581

38. A. Golbraikh, A. Tropsha, *J. Comput.-Aided Mol. Des.* **16** (2002) 357
39. A. Golbraikh, A. Tropsha, *J. Comput.-Aided Mol. Des.* **20** (2002) 269
40. A. Golbraikh, M. Shen, Z. Xiao, K. H. Lee, A. Tropsha, *J. Comput.-Aided Mol. Des.* **17** (2003) 241
41. F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, L. Eriksson, *J. Chemom.* **10** (1996) 521
42. P. P. Roy, S. Paul, I. Mitra, K. Roy, *Molecules* **14** (2009) 1660
43. Statistica 7.1, StatSoft Inc., Tulsa, OK, USA.