Use of Artificial Neural Networks in a QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives

M. Jalali-Heravi* and F. Parastar

Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran

Received July 22, 1999

Anti-HIV activity for a set of 107 inhibitors of the HIV-1 reverse transcriptase, derivatives of 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT), was modeled with the aid of chemometric techniques. The activity of these compounds was estimated by means of multiple linear regression (MLR) and artificial neural network (ANN) techniques and compared with the previous works. The results obtained using the MLR method indicate that the anti-HIV activity of the HEPT derivatives depends on the reverse of standard shadow area on the *YZ* plane and the ratio of the partial charges of the most positive atom to the most negative atom of the molecule. The best computational neural network model was a fully-connected, feed-forward method with a 6−6−1 architecture. The mean-square error for the prediction set using this network was 0.372 compared with 0.780 obtained using the MLR technique. Comparison of the quality of the ANN of this work with different MLR models shows that ANN has a better predictive power.

1. INTRODUCTION

Human immunodeficiency virus type 1 (HIV-1) is responsible for the condition known as AIDS (acquired immunodeficiency syndrome). HIV-1 is a retrovirus, i.e., an RNA virus that utilizes an enzyme known as DNA polymerase or reverse transcriptase (RT) to produce a DNA provirus that is able to insert itself into the host DNA. Because of the crucial role of RT to HIV replication, inhibitors of this enzyme are potential therapeutic agents in the battle against HIV.[1]

One class of RT inhibitors are the nucleoside analogues like 3′-azido-3′-deoxythymidine (AZT) and 2′,3′-dideoxyinosine (ddI).[2] These dideoxy compounds cause DNA chain termination when they are incorporated into a growing DNA strand. However, it is found that the treatment of some of these nucleoside inhibitors such as AZT is sometimes associated with considerable site effects such as bone marrow suppression.[3] Another class of HIV-RT inhibitors are non-nucleoside inhibitors (NNRTIs), which like the nucleoside analogues block reverse transcriptase but have a different mode of inhibition of viral replication. These inhibitors include TIBO, HEPT, Nevirapine, Pyridinone, BHAP, and α-APA.[4] Among them HEPT has proved to be a potent and selective inhibitor of HIV-1.[5,6] Other animal retroviruses and even HIV-2 are totally unaffected by this compound.[5]

Recently a growing number of critics question HIV's property and its relationship to the disease we term AIDS.[7] This ambiguity and designing new HEPT derivatives require a more detailed knowledge of the mechanism of RT inhibition by this class of compounds. The quantitative structure−activity relationships (QSAR) represent one of the most effective computational approaches for inspection of inhibition mechanism.[8−13]

Among different methods of chemometrics, multiple linear regression (MLR) and partial least squares (PLS) are two methods that have been used in QSAR study of RT inhibitors.[14] Kireev and co-workers have used MLR to relate the RT inhibitory activity of 87 analogues of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT).[9] Luco and Ferretti have developed a QSAR based on MLR and PLS methods for the anti-HIV activity of a large group of HEPT derivatives.[14] These researchers concluded that PLS is a better approach to MLR for improving the interpretability of the data and also for exhibiting models with a better predictive power. However, one problem in MLR and PLS methods is the fact that by using computational methods to generate descriptors, a data set may contain more descriptors than compounds. In this case the correlations observed may be chance correlations.[15] A second problem encountered in using regression analysis is the assumption of a linear relationship between the biological activity and one or more descriptors. On the other hand, biological phenomena are considered nonlinear by nature, and therefore the contribution of some of the parameters to RT inhibition properties can be nonlinear. The key of solving these problems is using neural networks, owing to their nonlinear mapping. As far as we are aware there is only one report on the use of artificial neural networks in designing HIV-1-RT inhibitors.[16] There is no emphasis on different parameters affecting RT inhibition properties and only the classification of 44 molecules such as AZT, dde, etc. is reported in this work.

In the present work both artificial neural network (ANN) and MLR techniques were used for modeling of the observed anti HIV-1 activity of 80 HEPT derivatives. These molecules represent most of the compounds of this class for which precise activity data are available. The adequacy of the developed QSAR models was examined by means of the prediction of anti HIV-1 activity of 27 HEPT derivatives for which imprecise activity data have been reported. Several descriptors were used in the characterization of the compounds. Also, the quality of QSAR, derived by means of

10.1021/ci990314+ CCC: $19.00
Published on Web 12/02/1999

ANN and MLR, has been compared with those obtained in the previous work.[14] It is shown here that neural networks were superior to regression methods in providing a good prediction of RT inhibitor activity of HEPT analogues.

2. METHODOLOGY

There are many types of network architectures, but the type that has been most useful for structure−activity relationships is the multilayer feed-forward (MLFF) network with back-propagation (BP) learning. This method was discovered independently by several researchers for different reasons.[17] The back-propagation learning method can be applied to any multilayer network that uses differentiable activation functions and supervised training. Like the δ rule, it is an optimization procedure based on gradient descent that adjusts weights to reduce the system error. During the learning phase in the BP method, input patterns are presented to the network in some sequence. Each training pattern is propagated forward layer by layer until an output pattern is computed. The computed output is then compared to a desired or target output, and an error value is determined. The errors are used as inputs to feedback connections from which adjustments are made to the synaptic weights layer by layer in a backward direction. The backward linkages are used only for the learning phase, whereas the forward connections are used for both the learning and the operational phases.[18]

In conventional BP, minimization is with respect to a defined mean-square error (MSE). The choice of MSE will depend on the way in which we wish to assess the performance, and, hence, how we should train the network. For the BP type of learning method, we do require that the error measure chosen be differentiable and tend to zero as the collective differences between the target and computed patterns ($t^p - O^p$) decrease over the entire training set. However, the MSE measure given by

$$E_{\text{tot}} = 1/P \sum_{P=1}^{P} E_P \tag{1}$$

where

$$E_P = 1/2 \sum_{k=1}^{m} (t_k^P - O_k^P)^2 \tag{2}$$

is by far the most popular error function. In this equation, P stands for the number of patterns and k is the number of neurons at the output layer. This function is often chosen because of its statistical properties and because it is better understood than other measures. It is a non-negative, differentiable function that penalizes large errors more than small ones.

3. EXPERIMENTAL SECTION

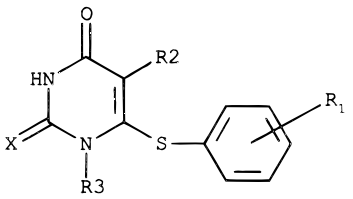
3.1. Biological Data. The activity data were taken from ref 14. The data set consists of 107 inhibitors of the HIV-1 reverse transcriptase, derivatives of 1-[2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT). This set was divided into a training set and a prediction set. The training set includes 80 compounds with precise activity data, and the prediction set includes 27 compounds whose imprecise activity data were reported. The chemical structures for the training and prediction sets are given in Table 1. The value of log $1/C$ was used as the dependent variable in which C represents the molar concentration of drug required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (HTLV-III_B strain).[19]

3.2. Descriptor Generation. The second step in developing the model was the numerical description of the molecular structures by defining descriptors. These descriptors were responsible for encoding important features of the structures and have been categorized as topological, geometric, electronic, and physicochemical properties. Topological descriptors include molecular connectivity and fragment descriptors.[20] Molecular connectivity descriptors encoded information about the type of atom, size, and degree of branching in the molecule. Fragment descriptors include the counts of atoms, bonds, rings, substructures, type of hybridation, etc. In order to calculate the geometric descriptors, the strain energy of the molecular structures must first be minimized. Therefore, the three-dimensional structure of each molecule was optimized using the semiempirical molecular orbital program of MOPAC with the AM1 Hamiltonian.[21] These descriptors include the moment of inertia, van der Waals molecular volume, surface of the molecule, shadow areas, and standard shadow areas.[22,23] Electronic descriptors encoded information about the electronic environment of each molecule, such as partial charges on the most negative and the most positive atoms and the ratio of them. Physicochemical descriptors include Hansch and Hammett constants.[24]

3.3. Regression Model. Because of the large number of descriptors considered, a stepwise multiple linear regression procedure based on the forward-selection and backward-elimination methods was used for inclusion or rejection of descriptors in the screened models.[25] In order to avoid overstimations or difficulties in interpretation of the resulting models, pairs of variables with an $r \geq 0.95$ were classified as intercorrelating ones, and only one of these was included in the screened model. A total of 59 descriptors were generated for each molecule. Fifteen parameters out of 59 were eliminated owing to intercorrelation, and hence 44 descriptors were submitted to the regression routine. Many models were generated by using this method. However, an ideal model is one that has high r and F values, low standard deviation, least numbers of independent variables, and high ability for prediction. The best MLR model is presented in Table 2.

3.4. Artificial Neural Network Generation. The ANN program was written in FORTRAN 90 in our laboratory. All of the calculations presented by authors were carried out on a Pentum II-MMX, 300 MHz. The network was generated by using the descriptors that have appeared in the MLR model as inputs. A three-layer network with a sigmoidal transfer function and δ rule learning was designed. Before the learning network was applied, the input vector and output values were normalized between 0.1 and 0.9. The transfer function has minimum and maximum values of 0 and 1, respectively. The normalizing of output values between 0.1 and 0.9 allows the network to slightly exceed the minimum and maximum values that were given in the original data

Table 1. Chemical Structures Along with the Observed and Calculated Values of Anti-HIV Activity for the HEPT Derivatives



						this work		previous work[a]	
no.	R1	R2	R3	X	obs	MLR	ANN	PLS	MLR
			Training Set						
1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	4.47	5.15	3.82	3.84
2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	3.76	3.83	4.18	4.10
3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	4.65	5.03	4.89	4.94
4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	4.83	4.97	5.43	5.44
5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	5.11	4.94	5.52	5.65
6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	5.36	4.97	4.98	4.93
7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	4.09	4.39	4.74	4.62
8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	4.54	4.80	5.33	5.29
9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	4.86	4.98	5.34	5.26
10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	5.08	4.99	5.34	5.24
11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.64	5.23	5.37	5.26
12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.62	4.33	4.76	4.57
13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	4.72	4.71	5.09	4.93
14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	5.50	5.07	5.23	5.23
15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	6.27	6.44	6.26	6.42
16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	6.25	6.21	6.20	6.28
17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	6.60	6.33	6.31	6.50
18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	5.37	4.83	4.80	4.63
19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	5.55	5.12	4.64	4.36
20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.27	5.08	4.93	4.72
21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.60	5.18	5.96	5.68
22	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	6.35	6.92	6.88	6.74
23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	6.79	5.88	6.17	6.01
24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.23	6.75	6.15	7.33	7.32
25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	7.48	7.69	7.82	7.76
26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.45	8.26	8.23	8.30
27	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	8.12	7.84	7.77	7.64
28	H	Et	$CH_2OCH_2CH_2CH_2OH$	O	6.92	5.97	6.85	6.84	6.66
29	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47	6.07	5.43	6.12	5.93
30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.20	6.79	6.83	7.28	7.24
31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	7.24	7.79	7.77	7.68
32	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.20	8.55	8.18	8.22
33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	7.51	7.84	7.72	7.56
34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66	5.36	3.71		5.39[b]
35	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	4.90	5.04	5.23	5.30
36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	5.26	5.27	5.28	5.38
37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	5.68	5.41	5.46	5.66
38	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	4.80	5.32	6.75	6.53
39	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.22	5.40	5.22	4.85	4.75
40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	5.28	5.08	5.04	4.81
41	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	6.07	4.92	6.10	5.70	6.06
42	H	Me	$CH_2OCH_2CH_2OMe$	O	5.06	5.26	5.14	5.18	5.35
43	H	Me	$CH_2OCH_2CH_2OAc$	O	5.17	5.24	5.05	4.56	4.62
44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12	5.53	5.24	5.59	5.66
45	H	Me	CH_2OCH_2Me	O	6.48	5.82	5.75	5.58	5.75
46	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.87	5.84	5.49	5.73
47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	6.07	5.18	4.74	4.74
48	H	Me	$CH_2OCH_2CH_2F$	O	5.96	5.49	5.39	5.19	5.23
49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.60	5.26	5.45	5.67
50	H	Me	CH_2OCH_2Ph	O	7.06	6.88	6.96	6.25	6.42
51	H	Et	CH_2OCH_2Me	O	7.72	6.72	6.80	7.20	7.14
52	H	Et	CH_2OCH_2Me	S	7.58	6.91	6.79	7.25	7.22
53	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	8.12	8.24	8.16	8.17
54	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.29	8.21	8.20	8.26
55	H	Et	CH_2OCH_2Ph	O	8.23	6.63	7.41	7.71	7.64
56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	8.54	8.36	8.50	8.51
57	H	Et	CH_2OCH_2Ph	S	8.09	7.53	8.09	7.76	7.72
58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	8.75	8.17	8.55	8.59
59	H	i-Pr	CH_2OCH_2Me	O	7.99	7.67	8.13	7.65	7.72
60	H	i-Pr	CH_2OCH_2Ph	O	8.51	8.53	8.19	8.10	8.16
61	H	i-Pr	CH_2OCH_2Me	S	7.89	7.84	7.79	7.70	7.80
62	H	i-Pr	CH_2OCH_2Ph	S	8.14	8.31	8.12	8.15	8.24

Table 1 (Continued)

						this work		previous work[a]	
no.	R1	R2	R3	X	obs	MLR	ANN	PLS	MLR
			Training Set						
63	H	Me	CH_2OMe	O	5.68	5.90	5.68	5.95	6.08
64	H	Me	CH_2OBu	O	5.33	5.73	5.64	5.31	5.58
65	H	Me	Et	O	5.66	6.29	5.51	6.45	6.64
66	H	Me	Bu	O	5.92	6.41	5.61	5.71	5.98
67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	8.00	7.81	8.15	8.13
68	H	Et	CH_2O-i-Pr	S	6.66	6.79	6.64	6.86	6.87
69	H	Et	CH_2O-c-Hex	S	5.79	6.63	5.95	5.98	6.06
70	H	Et	CH_2OCH_2-c-Hex	S	6.45	6.57	7.06	5.89	6.01
71	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	6.97	7.63	7.56	7.57
72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.44	7.99	7.59	7.63
73	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	6.61	6.82	7.59	7.60
74	3,5-Cl_2	Et	CH_2OCH_2Me	O	8.13	8.03	8.21	8.10	8.05
75	H	Et	CH_2O-i-Pr	O	6.47	6.79	6.71	6.81	6.78
76	H	Et	CH_2O-c-Hex	O	5.40	5.80	5.18	5.94	5.98
77	H	Et	CH_2OCH_2-c-Hex	O	6.35	6.44	6.45	5.84	5.93
78	H	Et	$CH_2OCH_2CH_2Ph$	O	7.02	7.10	7.16	7.55	7.52
79	H	c-Pr	CH_2OCH_2Me	S	7.02	7.39	6.85	6.79	6.61
80	H	c-Pr	CH_2OCH_2Me	O	7.00	7.16	7.01	6.74	6.53
			Prediction Set[c]						
81	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$-n	O	<4.46	6.01	5.24		5.12
82	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	<3.89	4.64	5.14		4.31
83	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	<3.53	4.39	4.83		4.60
84	4-F	Me	$CH_2OCH_2CH_2OH$	O	<3.60	4.50	3.58		5.10
85	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	<3.60	5.04	3.80		5.45
86	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	<3.72	4.43	4.22		5.04
87	4-CN	Me	$CH_2OCH_2CH_2OH$	O	<3.60	5.65	4.17		4.98
88	4-OH	Me	$CH_2OCH_2CH_2OH$	O	<3.56	4.95	3.57		4.74
89	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	<3.60	4.98	3.61		5.28
90	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	<3.96	4.91	3.68		5.00
91	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	<3.45	4.69	4.56		4.67
92	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	<3.51	4.99	4.68		4.14
93	H	COOMe	$CH_2OCH_2CH_2OH$	O	<5.18	4.22	5.01		5.28
94	H	CONHPh	$CH_2OCH_2CH_2OH$	O	<4.74	5.29	5.13		4.60
95	H	SPh	$CH_2OCH_2CH_2OH$	O	<4.68	5.01	5.57		4.97
96	H	C≡CH	$CH_2OCH_2CH_2OH$	O	<4.74	5.95	5.23		6.30
97	H	C≡CPh	$CH_2OCH_2CH_2OH$	O	<5.47	5.86	5.26		4.71
98	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	<3.60	4.76	5.02		4.57
99	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	<4.92	5.78	4.58		5.27
100	H	COPh	$CH_2OCH_2CH_2OH$	O	<4.89	5.08	5.15		5.11
101	H	C≡CMe	$CH_2OCH_2CH_2OH$	O	<4.72	6.09	5.28		5.57
102	H	F	$CH_2OCH_2CH_2OH$	O	<4.00	5.25	5.11		5.00
103	H	Cl	$CH_2OCH_2CH_2OH$	O	<4.52	5.46	5.16		5.36
104	H	Br	$CH_2OCH_2CH_2OH$	O	<4.70	5.49	5.16		5.47
105	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	<4.70	5.61	5.63		5.93
106	H	Me	H	O	<3.60	6.06	5.52		5.63
107	H	Me	Me	O	<3.82	6.16	5.49		5.75

[a] The calculated values of training set were taken from ref 14. [b] This molecule was absent in the training set and is calculated using Luco's model.[14] [c] Last column consists of the values that have been calculated using Luco's equation.[14]

Table 2. Best MLR Model for the Prediction of Anti-HIV Activity[a]

descriptor	notation	regression coefficient	mean effect
(1) reciprocal of the standard shadow area on *YZ* plane	$1/S$	$3.307(\pm0.679)$	5.75
(2) ratio of the partial charges on the most positive and the most negative atoms	POS/NEG	$2.013(\pm0.741)$	−2.19
(3) heat of formation (kcal/mol)	ΔH_f	$6.689 \times 10^{-3}(\pm1.631 \times 10^{-3})$	−0.51
(4) square of the number of SP^3 carbon atoms of the R2 substituent	$(NCSP3\text{-}R2)^2$	$0.141(\pm0.028)$	0.44
(5) number of hydroxyl groups on the R3 substituent	NOH-R3	$-0.804(\pm0.171)$	−0.41
(6) cub of summation of the positions of R1 on the C-6 aromatic ring	$(NS\text{-}R1)^3$	$5.097 \times 10^{-3}(\pm9.938 \times 10^{-4})$	0.24
constant		$2.977(\pm1.335)$	

[a] The statistics for this model are $n = 80$, $r = 0.901$, SE = 0.607, and F = 52.

file. The initial weight was selected randomly between −1 and 1. The number of neurons in the hidden layer, momentum, and learning rate were optimized in this work. At this point, the MSE was plotted versus the number of neurons at the hidden layer at the arbitrary learning rate and momentum and 10 000 iterations (Figure 1a). The number of neurons at the hidden layer at the minimum of this curve was selected as the optimum number. After this step the learning rate was varied from 0.1 to 0.9, and for each learning rate the momentum was examined from 0.1 to 0.9 (Figure 1b). A

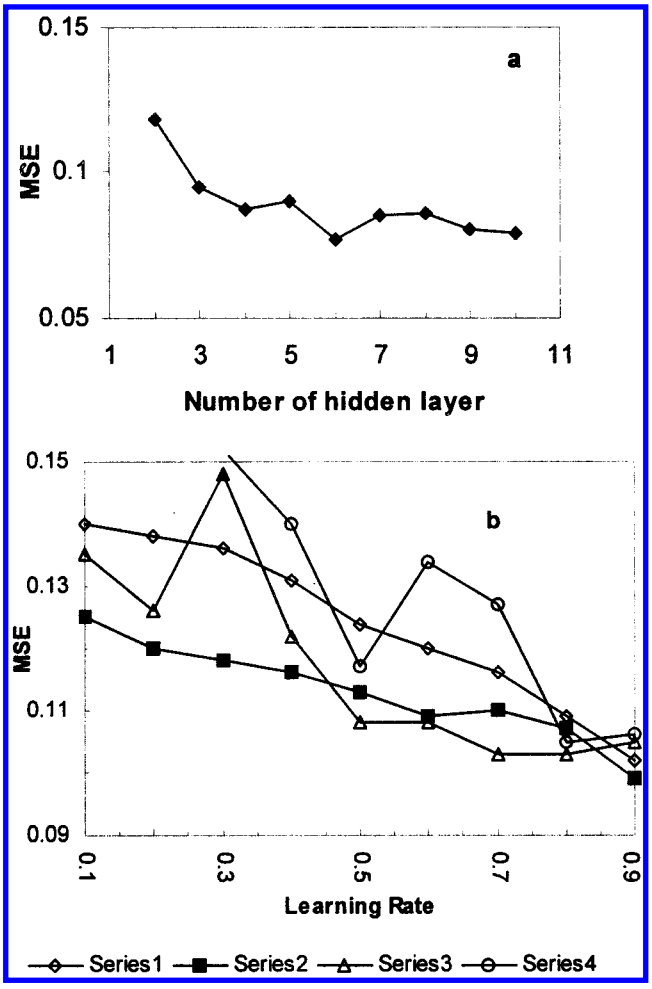


Figure 1. Variations of MSE versus (a) number of hidden layer and (b) learning rate for different momentums.

total of 81 networks were designed in this way. Finally, the number of the neurons at the hidden layer with the use of optimized momentum and learning rate was determined.

4. RESULTS AND DISCUSSION

The chemical structure and the observed and calculated values of the activity for the compounds studied in this work are given in Table 1. The activities calculated by Luco and Ferretti are also included in this table for comparison.[14]

4.1. Multiple Regression Analysis. MLR was performed on the compounds described in Table 1. We have included all 80 molecules of the training set for the model generation, while Luco and Ferretti have excluded compound 34 for their equation.[14] The reason for this was the fact that none of the descriptors used in their study can account for the complete loss of activity that results from the introduction of substituents at the 4-position of the 6-(phenylthio) moiety of HEPT.[14] After screening of the descriptors and submission of 44 parameters to the regression routine, a few suitable models were obtained, and among them the best model was selected and presented in Table 2. It can be seen from this table that six descriptors have appeared in the model. These descriptors consist of 1/*S*, which is a geometric descriptor; POS/NEG and ΔH_f, which can be considered as electronic descriptors; and NS-R1, NCSP3-R2, and NOH-R3 parameters, which are topological descriptors. The parameter 1/*S* shows a mean effect of 5.75, which is the largest one between the other mean effects in Table 2. The *S* in this parameter stands for the standard shadow area on the *YZ* plane. The molecules are drawn in such a way that two rings of the HETPs are in the *XY* plane. Therefore, for the molecules whose substituents are large and can be positioned out of the *XY* plane, the shadow area on the *YZ* plane is large and hence shows a smaller activity. This is in agreement with the experiment. As can be seen from Table 1, the molecules with larger R3 substituents show generally less anti-HIV activity compared with those with smaller substituents. The electronic descriptor of POS/NEG also shows a considerable but negative mean effect. This reveals that in addition to the geometry of the HEPTs, which plays a major role in anti-HIV activity, the electronic properties of these molecules are also important. This is also in agreement with the previous efforts indicating that changing the size or attaching polar groups to the substituents of the HEPTs decreases the activity.[19,26−28] The presence of NS-R1, NCSP3-R2, and NOH-R3 in the selected model reveals that all three substituents play some roles in the anti-HIV activity of HEPT derivatives. The calculated values of all descriptors appearing in the best model are given in Table 3. It is noteworthy that there is no significant intercorrelation between the descriptors appearing in the selected model. It should be noted that all of these parameters are calculated values while some of the descriptors appearing in the equation derived by Luco and Ferretti[14] should be obtained by the experiment. We have used two strategies for testing the validity of the selected MLR model. In the first strategy a cross-validation method[29] was used for which the *Q* coefficient is the cross-validated r^2 that describes the predictive power of the model. The average *Q* value in our case was 0.778, which should be compared with the 0.826 obtained by Luco and Ferretti.[14] However, the better value obtained by these authors may be due to the leave-one-out procedure that they have used in their work.[14] In the present work, since six descriptors have appeared in the model then six molecules have been removed randomly from the training set each time and a model was developed with the remaining molecules. Then, the anti-HIV activities of the six molecules were predicted by this equation. This process was repeated until each molecule had a chance to be predicted once. The results of the cross-validation method are given in Table 4. As a second strategy for the evaluation of the developed QSAR model, the anti-HIV activity of 27 HEPT derivatives was predicted by using the selected MLR model (compounds 81−107, Table 1). It seems appropriate to compare the results obtained by using equation 4 given in the previous work[14] with the results presented in this study. The calculated MSEs for the prediction set were 0.780 and 0.676 for the MLR equations generated in this and previous works, respectively. It should be mentioned that Luco and co-workers have included a descriptor called I-4R1 in their model.[14] This parameter takes the value 1 or 0 for the presence or absence of a substituent at the 4-position of the C-6 aromatic ring. Inclusion of this descriptor improved our model considerably. However this descriptor has only one nonzero value in our and Luco's training set, and from a statistical point of view it cannot be included in the model. On the other hand, a lower number of descriptors (six parameters) have appeared in our model compared with the previous MLR model (nine descriptors).[14]

Table 3. Calculated Values of the Descriptors Appearing in the Best MLR Model

no.[a]	1/*S*	POS/ NEG	ΔH_f	(NCSP3-R2)^2	(NS-R1)^3	NOH-R3	no.[a]	1/*S*	POS/ NEG	ΔH_f	(NCSP3-R2)^2	(NS-R1)^3	NOH-R3
1	1.5521	−1.1100	−113.09	1	8	1	55	1.6337	−1.0273	−36.14	4	0	0
2	1.5716	−1.5408	−100.35	1	8	1	56	1.9238	−1.0470	−51.03	4	216	0
3	1.6480	−1.0949	−138.05	1	8	1	57	1.7995	−1.0474	21.07	4	0	0
4	1.6337	−1.1152	−113.45	1	27	1	58	1.8921	−1.0916	7.18	4	216	0
5	1.7138	−1.0860	−118.78	1	27	1	59	1.8172	−1.0357	−71.98	9	0	0
6	1.7889	−1.0570	−128.11	1	27	1	60	2.0044	−1.0337	−37.80	9	0	0
7	1.8132	−1.2835	−257.79	1	27	1	61	1.8723	−1.2377	−15.12	9	0	0
8	1.6044	−1.0733	−153.50	1	27	1	62	1.9216	−1.2063	20.09	9	0	0
9	1.6239	−1.0905	−111.42	1	27	1	63	1.6051	−1.0510	−59.51	1	0	0
10	1.6639	−1.0726	−103.56	1	27	1	64	1.5903	−1.0448	−78.60	1	0	0
11	1.6703	−1.0498	−32.24	1	27	1	65	1.6711	−1.0883	−23.13	1	0	0
12	1.8129	−1.5653	−97.88	1	27	1	66	1.7343	−1.0868	−36.84	1	0	0
13	1.6410	−1.0314	−156.69	1	27	1	67	1.8119	−1.2229	−25.28	4	216	0
14	1.8854	−1.0865	−145.53	1	27	1	68	1.6941	−1.1341	−10.77	4	0	0
15	1.7937	−1.1118	−121.20	1	216	1	69	1.7082	−1.1695	−30.25	4	0	0
16	1.7841	−1.1231	−116.10	1	216	1	70	1.6305	−1.0510	−36.15	4	0	0
17	1.7355	−1.0454	−65.59	1	216	1	71	1.6521	−1.0587	14.90	4	0	0
18	1.9320	−1.0667	−191.59	1	27	1	72	1.7912	−1.0538	13.62	4	0	0
19	1.8893	−1.0667	−145.66	1	27	1	73	1.5385	−1.0524	14.06	4	0	0
20	1.6653	−1.0695	−77.31	1	27	1	74	1.8409	−1.0577	−82.80	4	216	0
21	1.8332	−1.0784	−88.50	1	0	1	75	1.7781	−1.0536	−74.38	4	0	0
22	1.8911	−1.1208	−55.40	4	0	1	76	1.5354	−1.1133	−83.87	4	0	0
23	1.7765	−1.0353	−63.29	9	0	1	77	1.7135	−1.0532	−94.10	4	0	0
24	1.7409	−1.0161	−57.46	9	0	1	78	1.8047	−1.0472	−43.42	4	0	0
25	1.9410	−1.1342	−71.93	4	216	1	79	1.5434	−1.0597	25.11	9	0	0
26	1.9459	−1.0077	−72.20	9	216	1	80	1.5878	−1.0439	−33.26	9	0	0
27	2.0412	−1.0730	−45.20	4	216	1	81	1.7940	−1.0660	−130.55	1	0	0
28	1.8636	−1.0631	−114.45	4	0	1	82	1.5765	−1.0598	−114.63	1	8	1
29	1.6932	−1.0609	−121.26	9	0	1	83	1.5793	−1.1008	−155.73	1	27	1
30	1.9084	−1.0713	−116.41	9	0	1	84	1.5366	−1.0717	−154.20	1	64	1
31	1.9414	−1.0557	−129.71	4	216	1	85	1.6179	−1.0692	−115.70	1	64	1
32	2.0125	−1.0434	−129.55	9	216	1	86	1.7135	−1.5648	−104.80	1	64	1
33	2.0367	−1.0905	−126.49	4	216	1	87	1.7283	−1.0800	−76.67	1	64	1
34	1.7056	−1.0514	−117.25	1	64	1	88	1.6807	−1.1031	−148.42	1	64	1
35	1.6790	−1.1192	−104.01	1	0	1	89	1.6824	−1.1077	−143.98	1	64	1
36	1.6279	−1.0333	−51.57	1	0	1	90	1.6622	−1.1213	−140.23	1	64	1
37	1.7771	−0.8831	−85.65	0	0	1	91	1.7677	−1.1281	−193.97	1	27	1
38	1.6295	−1.0908	−82.68	0	0	1	92	1.6345	−0.9326	−142.88	1	27	1
39	1.7522	−1.0644	−61.64	0	0	1	93	1.6067	−1.0866	−177.38	1	0	1
40	1.7259	−1.0913	−79.43	1	0	1	94	1.8096	−1.0840	−99.11	0	0	1
41	1.5337	−1.0547	−29.55	0	0	1	95	1.5547	−0.9164	−66.97	0	0	1
42	1.5430	−1.1263	−99.95	1	0	0	96	1.8947	−1.0856	−44.31	0	0	1
43	1.5860	−1.0464	−147.40	1	0	0	97	1.8155.	−1.0701	−23.39	0	0	1
44	1.6062	−1.0611	−111.23	1	0	0	98	1.5957	−1.1131	−105.18	1	27	1
45	1.5810	−1.0378	−65.59	I	0	0	99	1.6776	−1.1048	−142.56	9	0	1
46	1.6116	−1.0294	−72.72	1	0	0	100	1.7397	−1.0806	−97.92	0	0	1
47	1.4741	−1.0331	22.35	I	0	0	101	1.8947	−1.0856	−44.38	1	0	1
48	1.5878	−1.0477	−112.37	1	0	0	102	1.8904	−1.0934	−141.04	0	0	1
49	1.5674	−1.1141	−66.26	1	0	·0	103	1.8723	−1.0820	−105.21	0	0	1
50	1.8416	−1.0537	−30.78	1	0	0	104	1.8584	−1.0835	−93.58	0	0	1
51	1.7504	−1.0572	−70.02	4	0	0	105	1.5475	−1.0491	−74.20	1	0	0
52	1.7188	−1.1415	−3.02	4	0	0	106	1.5803	−1.0879	−13.21	1	0	0
53	1.8639	−1.0435	−84.67	4	216	0	107	1.6192	−1.0923	−17.18	1	0	0
54	1.8653	−1.1550	−28.16	4	216	0							

[a] Numbers refer-to the molecules given in Table 1.

4.2. Artificial Neural Network Analysis. The main goal of the present work was development of an ANN to predict the anti-HIV activity for the HETP derivatives. Therefore, an ANN was generated by using the descriptors appearing in the MLR model as inputs. However, using the same descriptors make the comparison of two models possible. A 6−6−1 neural network was developed with the optimum momentum and learning rate of 0.9 and 0.2, respectively. In order to prevent the overfitting, the calculated MSEs for the training and prediction sets were plotted against the number of iterations (Figure 2). The overfitting will start after 18 000 trainings of the network. The ANN calculated values of the biological activity for the training and prediction sets are included in Table 1.

To evaluate the neural network, the MSE of its results is compared with the MSE of the regression model developed in our laboratory as well as with the previous one.[14] The MSEs were 0.170, 0.084, and 0.073 for the training set in the present MLR, Luco and Ferretti's model, and the ANN, respectively. The corresponding MSEs for the prediction set were 0.780, 0.676, and 0.372, respectively. This reveals the superiority of the neural network over that of the regression models. It should be noted that the descriptors used as inputs in the ANN are those chosen by the regression analysis

Table 4. Results of Cross-Validation Procedure

	Q^2 (n^a)	press		Q^2 (n^a)	press
1	0.926(6)	0.85	8	0.701(6)	2.59
2	0.904(6)	0.91	9	0.887(6)	0.98
3	0.765(6)	2.42	10	0.682(6)	2.93
4	0.766(6)	2.38	11	0.808(6)	1.86
5	0.571(6)	4.63	12	0.770(6)	3.81
6	0.772(6)	2.17	13	0.856(6)	1.49
7	0.954(6)	0.43	14	0.525(2)	3.36

[a] Numbers in parentheses represent the number of the molecules that have been removed each time in model development and were predicted by the generated model.

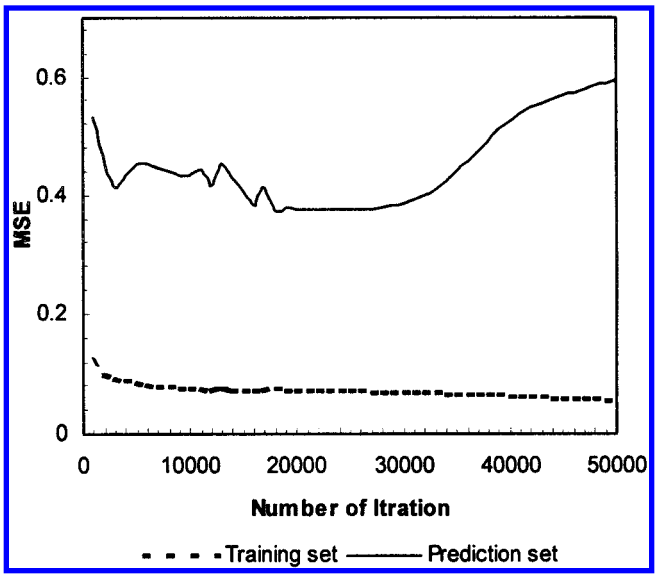


Figure 2. Variations of MSE versus the number of iterations for the training and prediction sets.

procedure. Therefore, the superiority of the ANN over that of the MLR is partly due to the fact that in ANNs the interactions between different parameters used are considered, whereas for the MLR techniques these parameters should be independent. In order to verify the descriptors introduced by Luco and Ferretti, we have developed a neural network with their descriptors. However, this model started overfitting after 7000 iterations. This means that their descriptors are not able to show all features of the inhibitory property of the HEPT derivatives. The observed values for the log 1/*C* of the molecules in the prediction set are the high limit of this quality, and in fact the real values are less than that of those given in Table 1. It can be seen from this table that the ANN prediction is correct for five molecules of the prediction set. This should be compared with one and two correct predictions for our and previous MLR models. For most of the molecules given in Table 1 the calculated values of the anti-HIV activity obtained using all three models are higher than that of the observed high limits. This indicates that the experimental values for these compounds may be underestimated. Figure 3 shows the plot of calculated against the experimental anti-HIV activity for the HEPT derivatives together with their correlation coefficient. The residuals of ANN predicted values of activities are plotted against the experimental values in Figure 4. The propagation of residuals on both sides of zero indicates that no systematic error exists in the development of the neural network.

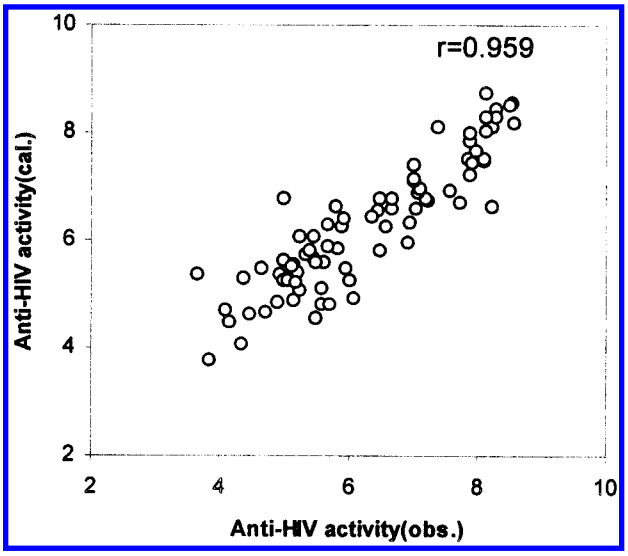


Figure 3. Plot of calculated log 1/*C* against the experimental values.

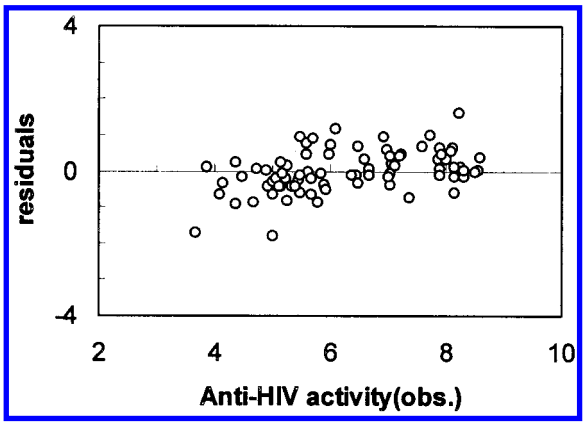


Figure 4. Plot of residuals versus experimental values of anti-HIV activity.

From the superiority of ANN over that of the MLR methods, one may conclude that the contribution of some of the parameters to RT inhibition property can be nonlinear. This conclusion arises from the fact that the same descriptors have been used for the development of the MLR model and the artificial neural network.

REFERENCES AND NOTES

(1) Jacobo-Molina, A.; Ding, J.; Nanni, R. G.; Clark, A. D.; Lu, X.; Tantillo, C.; Williams, R. L.; Kamer, G.; Ferris, A. L.; Clerck, P.; Hizi, A.; Hughes, S.; Arnold, E. Crystal Structure of Human Immunodeficiency Virus Type 1 Reverse Transcriptase Complexed with Double-Stranded DNA at 3.0 Å Resolution Shows Bent DNA. *Proc. Natl. Acad. Sci.* **1993**, *90*, 6320−6324.

(2) Schinazi, R. F. Competitive Inhibitors of Human Immunodeficiency Virus Reverse Transcriptase. *Perspect. Drug Disc. Des.* **1993**, *1*, 151−180.

(3) INFO Pool: Anti HIV Treatment. http://www.med.ed.ac.uk/ridu.

(4) De Clercq, E. HIV 1 Specific RT Inhibitors: Highly Selective Inhibitors of Human Immunodeficiency Virus Type 1 That Are Specifically Targeted at the Viral Reverse Transcriptase. *Med. Res. Rev.* **1993**, *13*, 229−258.

(5) Baba, M.; Tanaka, H.; De Clercq, E.; Pauwels; R.; Balzarini, J.; Schols, D.; Nakashima, H.; Perno, C. F.; Wolker, R. T.; Miyasaka, T. Highly Specific Inhibitation of Human Immunodeficiency Virus Type 1 by a Novel 6-Substituted Acyclouridine Derivative. *Biochem. Biophys. Res. Commun.* **1989**, *165*, 1375−1381.

(6) Miyasaka, T.; Tanaka, H.; Baba, M.; Hayaakawa, H.; Walker, R. T.; Balzarini, J.; De Clercq, E. A Novel Lead for Specific Anti-HIV-1

Agents: 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J Med. Chem.* **1989**, *32*, 2507−2509.
(7) Root-Bernstein, R. Aids Dissidents Speak Out. http://www.garvnull.com/Documents/aids.htm.
(8) Hansch, C.; Zhang, L. QSAR of HIV Inhibitors. *Bioorg. Med. Chem. Lett.* **1992**, *2*, 1165−1169.
(9) Kireev, D. B.; Chrétien, J. R.; Grierson, D. S.; Monneret, C. A 3D QSAR Study of a Series of HEPT Analogous: The Influence of Conformational Mobility on HIV-1 Reverse Transcriptase Inhibitation. *J. Med. Chem.* **1997**, *40*, 4257−4264.
(10) Gussio, R.; Pattabiraman, N.; Zaharevitz, D. W.; Kellogg, G. E.; Topol, I. A.; Rice, W. G.; Schaeffer, C. A.; Erickson, J. W.; Burt, S. K. All-Atom Models for the Non-Nucleoside Binding Site of HIV-1 Reverse Transcriptase Complexed with Inhibitors: A 3D QSAR Approach. *J. Med. Chem.* **1996**, *39*, 1645−1650.
(11) Hannongbua, S.; Lawtrakul, L.; Limtrakul, J. Structure-Activity Correlation Study of HIV-1 Inhibitors. Electron and Molecular Parameters. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 145−152.
(12) Hannongbua, S.; Lawtrakul, L.; Sotriffer, C. A.; Rode, B. M. Comparative Molecular Field Analysis of HIV-1 Reverse Transcriptase in Inhibitors in the Class of 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *Quant. Struct.-Act. Relat.* **1996**, *15*, 389−394.
(13) Bacheler, L. T.; Rayner, M. M.; Erickson-vitanen, J. L.; Jackson, D. A.; Calabrese, C.; Schadt, M.; Chang, C. H. Nonpeptide Cyclic Cyanoguanidines as HIV-1 Protease Inhibitors: Synthesis, Structure−Activity Relationships, and X-ray Crystal Structure Studies. *J. Med. Chem.* **1998**, *41*, 1446−1455.
(14) Luco, J. M.; Ferretti, 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.* **1997**, *37*, 392−401.
(15) Wikel, J. H.; Dow, E. R.; Heathman, M. Interpretative Neural Networks for QSAR. *Net. Sci.* http://www.netsci.org/Science/Combichem/feature02.html.
(16) Tetko, I. V.; Tanchuk, V. Y.; Chentsova, N. P.; Antonenko, S. V.; Poda, G. I.; Kukhar, V. P.; Luik, A. I. HIV-1 Reverse Transcriptase Inhibitor Design using Artificial Neural Networks. *J. Med. Chem.* **1994**, *37*, 2520−2526.
(17) Werbos,V. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. Thesis, Harvard University, 1994.
(18) Patterson, D. W. *Artificial Neural Networks: Theory andApplications*; Simson & Schuster: New York, 1996; Part III, Chapter 6.
(19) 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.* **1992**, *35*, 4713−4719.
(20) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley and Sons: New York, 1986.
(21) *MOPAC package*, Version 6; U. S. Air Force Academy, Colorado Springs, CO 80840.
(22) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(23) Rohrbaugh, R. H.; Jurs, P. C. Descriptors of Molecular Shape Applied in Studies of Structure−Activity.and Structure−Property. *Anal. Chim. Acta.* **1987**, *199*, 99−109.
(24) Hansch, C.; Leo, A.; Hoekman, D. In *Hydrophobic, Electronic and Steric Constants*; Exploring QSAR, Vol. 2; Heller, S. R., Ed.; American Chemical Society: Washington D.C., 1995; pp 217−304.
(25) SPSS for Windows, Statistical Package for IBM PC, SPSS Inc. 1993. http://www.spss.com.
(26) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Inouye, N.; Baba, M.; Shigeta, Sh.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of Deoxy Analogs of 1-[2-Hydroxy-ethoxy)methyl]-6-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1995**, *38*, 2860−2865.
(27) Tanaka, H.; Baba, M.; Sakamaki, H.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, Sh.; Walker, R. T.; Balzarini, J.; De Clercq, E. A New Class of HIV-1 Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV-1 Activity 5- or 6-Substituted Analogues of 1-[2-Hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349−357.
(28) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, Sh.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Structure−Activity Relationships of 1-[2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine Analogues: Effect of Substitutions at the C-6 Phenyl Ring and at the C-5 Position on Anti-HIV-1 Activity. *J. Med. Chem.* **1992**, *35*, 337−345.
(29) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.

CI990314+