Neural Networks: Accurate Nonlinear QSAR Model for HEPT Derivatives

Latifa Douali,*,† Didier Villemin,‡ and Driss Cherqaoui†

Département de Chimie, Faculté des Sciences Semlalia BP 2390 Université Cadi Ayyad, Marrakech, Morocco, and Ecole Nationale Supérieure d'Ingénieurs (E.N.S.I.) I.S.M.R.A., LCMT, UMR CNRS no. 6507, 6 boulevard Maréchal Juin, 14050 Caen Cedex, France

Received March 8, 2003

A nonlinear quantitative structure−anti-HIV-1-activity relationship (QSAR) study was investigated in a series of 1-[2-hydroxyethoxy-methyl]-6-(phenylthio) thymine] (HEPT) derivatives acting as nonnucleoside reverse transcriptase inhibitors (NNRTIs). This QSAR study has been undertaken by a three-layered neural network (NN) using molecular descriptors known to be responsible for the anti-HIV-1 activity. The usefulness of the model and the nonlinearity of the relationship between molecular descriptors and anti-HIV-1 activity have been clearly demonstrated. The obtained model outperforms those given in the literature in both the fitting and predictive stages. NN analysis yielded predicted activities in excellent agreement with the experimentally obtained values ($R^2 = 0.977$, predictive $r^2 = 0.862$). The effect of each molecular feature on the anti-HIV-1 activity variation has been clearly elucidated.

1. INTRODUCTION

Because human type 1 immunodeficiency virus (HIV-1) is the causative agent of AIDS, extensive works are currently going on to block its replication. Reverse transcriptase (RT) is an attractive target for chemotherapeutic intervention,[1] so it should be possible to identify drugs that specifically interfere with RT but have little or no effect on the host.[2] Nonnucleoside reverse transcriptase inhibitors (NNRTIs) are anti-HIV-1-RT specific compunds that exhibit low cytotoxicity and produce few side effects. All the compounds reported as NNRTIs share a remarkable common feature. They consist of aromatic moieties and bind to a hydrophobic pocket near the polymerase catalytic site.[3] The binding event alters the conformation of critical residues and thereby inhibits the ability of the enzyme to perform normal RT functions.[4]

1-[2-hydroxyethoxy-methyl]-6-(phenylthio) thymine] (HEPT) is the first NNRTI analogue shown to have potent anti-HIV-1 activity and to inhibit HIV-1 at nanomolar concentration.[5] Since this discovery, the synthesis of a large number of HEPT derivatives has been performed by the group of Tanaka et al.[6−9] and they have been extensively studied for many years.[3]

Although NNRTIs dramatically decrease the viral load in most infected persons upon initiation of antiviral therapy, resistance to the chemotherapeutics usually develops.[3] The use of NNRTIs in combination therapy with NRTIs and with protease inhibitors is currently the best method for controlling HIV infections, but there is a further need for the development of NNRTIs and the design of new and more effective drugs.

Fortunately, the availability of computational techniques based on quantitative structure−activity relationships (QSAR) has accelerated the drug design process. In fact, QSAR can be viewed as a technique attempting to summarize chemical and biological information in a form that allows one to generate relationships between chemical structure and biological activity.[10]

A certain number of computational techniques have been found useful for the establishment of these relationships such as multiple linear regression (MLR), principle component analysis (PCA), and partial least squares (PLS). However, in many cases these methods exhibit some limitations and give poor statistical results, especially when the relationship between the dependent and independent variables is so complex that it can not be emulated by a simple linear relationship. Recently, there is growing interest in the application of neural networks (NNs) in the field of QSAR. The special interest in NNs arises from their ability to perform nonlinear mapping of the physicochemical descriptors to the corresponding biological activity implicitly. It has been demonstrated that the NN technique is often superior to traditional techniques[11−13] and can overcome some limitations of these QSAR approaches.[14]

Whether by traditional methods or by NNs, the success of a QSAR study depends also on the selection of variables[15] (molecular descriptors) and on the representation of the information. Variables should give the maximum of information in the activity variations and collinearity among them must be kept to a minimum.

Recently, several QSAR studies on HEPT derivatives based on MLR, PLS, and NNs have been investigated.[16−21] The overall picture, which emerges from these QSAR studies, shows that the hydrophobic, electronic, and steric characteristics of the compounds have predominant roles in the anti-HIV-1 activity of HEPT derivatives. However, problems of the choice of molecular descriptors, the information they contain, and whether quadratic, cubic, or cross product terms should be included usually obscure the picture.

In a previous study,[14] we have clearly pointed out the nonlinear aspect of the structure−anti-HIV-1 activity rela-

* Corresponding author phone: (+212) 4437 65 59; fax: (+212) 4443 74 08; e-mail: l.douali@ucam.ac.ma or d.cherkaoui@menara.ma.
† Université Cadi Ayyad.
‡ Ecole Nationale Supérieure d'Ingénieurs.

10.1021/ci034047q CCC: $25.00
Published on Web 06/12/2003

Table 1. Chemical Structures of the Compounds Studied and Their Anti-HIV-1 Activity (log ($1/EC_{50}$))

	substituents				log($1/EC_{50}$)		
no.	R_1	R_2	R_3	*X*	obsd	calcd[a]	diff
1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	4.15	0.00
2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	3.87	0.02
3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	4.72	0.00
4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	5.47	−0.12
5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	5.48	−0.19
6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	4.95	0.03
7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	4.29	−0.06
8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	5.60	0.12
9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	5.15	0.26
10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	5.04	−0.20
11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.00	0.00
12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.93	0.46
13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	4.19	0.10
14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	4.91	0.25
15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	6.53	−0.06
16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	5.74	−0.15
17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	6.61	−0.05
18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	4.97	−0.13
19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	4.92	−0.21
20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	4.67	−0.33
21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.46	−0.14
22	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	7.07	0.11
23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	5.25	0.25
24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.23	7.11	−0.12
25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	8.11	0.00
26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.23	0.07
27	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	7.66	0.29
28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92	6.75	−0.17
29	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47	5.60	0.13
30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.20	7.43	0.23
31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	8.05	0.16
32	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.26	−0.30
33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	7.65	−0.20
34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66	4.18	0.52
35	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	5.15	0.00
36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	5.67	−0.34
37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	5.20	−0.24
38	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	5.69	0.00
39	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.22	5.24	0.02
40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	3.97	−0.40
41	H	Me	CH2OCH2CH_2OMe	O	5.06	5.02	−0.04
42	H	Me	CH2OCH2CH_2OAc	O	5.17	5.16	−0.01
43	H	Me	CH2OCH2CH_2OCHOPh	O	5.12	5.25	0.13
44	H	Me	CH_2OCH_2Me	O	6.48	6.73	0.25
45	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.59	−0.23
46	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	5.12	-0.12
47	H	Me	$CH_2OCH_2CH_2F$	O	5.96	6.01	0.05
48	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.41	−0.07
49	H	Me	CH_2OCH_2Ph	O	7.06	7.01	−0.05
50	H	Et	CH_2OCH_2Me	O	7.72	7.61	−0.11
51	H	Et	CH_2OCH_2Me	S	7.58	7.55	−0.03
52	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	8.42	0.18
53	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.36	0.06
54	H	Et	CH_2OCH_2Ph	O	8.23	7.72	−0.51
55	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	8.31	−0.24
56	H	Et	CH_2OCH_2Ph	S	8.09	8.14	0.05
57	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	8.35	0.21
58	H	i-Pr	CH_2OCH_2Me	O	7.99	7.74	−0.25
59	H	i-Pr	CH_2OCH_2Ph	O	8.51	8.28	−0.23
60	H	i-Pr	CH_2OCH_2Me	S	7.89	7.94	0.05
61	H	i-Pr	CH_2OCH_2Ph	S	8.14	8.18	0.04
62	H	Me	CH_2OMe	O	5.68	5.88	0.20
63	H	Me	CH_2OBu	O	5.33	5.46	0.13
64	H	Me	Et	O	5.66	5.64	−0.02
65	H	Me	Bu	O	5.92	5.93	0.01
66	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	8.16	0.27
67	H	Et	CH_2O−i-Pr	S	6.66	6.60	−0.06
68	H	Et	CH_2O−c-Hex	S	5.79	5.96	0.17
69	H	Et	CH_2OCH_2-c-Hex	S	6.45	6.47	0.02
70	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	7.24	0.13
71	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.55	−0.37
72	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	7.04	0.00

Table 1 (Continued)

	substituents				log(1/EC$_{50}$)		
no.	R_1	R_2	R_3	X	obsd	calcd[a]	diff
73	3,5-Cl_2	Et	CH_2OCH_2Me	O	8.13	8.18	0.05
74	H	Et	CH_2O−i-Pr	O	6.47	6.42	−0.05
75	H	Et	CH_2O−c-Hex	O	5.40	5.36	−0.04
76	H	Et	CH_2OCH_2-c-Hex	O	6.35	6.21	−0.14
77	H	Et	$CH_2OCH_2CH_2Ph$	O	7.02	7.68	0.66
78	H	c-Pr	CH_2OCH_2Me	S	7.02	7.16	0.14
79	H	c-Pr	CH_2OCH_2Me	O	7.00	7.24	0.24

[a] The values given are the outputs of the NN architecture 7−5−1.

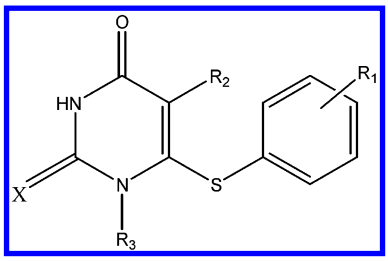

Figure 1. General structure of HEPT derivatives studied.

tionship of some HEPT derivatives by NN. We propose, then, to use this promising technique to develop a QSAR model with a better predictive power using descriptors known to be responsible for the anti-HIV-1 activity of these compounds. All the above-mentioned QSAR models are considered. Results will be analyzed to rapidly provide medicinal chemists with immediately useful features derived by NN analysis allowing the control of anti-HIV-1 activity of HEPT derivatives.

The goals of the current work are (i) to extract the relevant descriptors to establish the structure−activity relationship of a large set of HEPT derivatives; (ii) to bring up the high predictive power of NNs on the series of HEPT derivatives; and (iii) to elucidate the effect of the molecular properties on the anti-HIV-1 activity.

2. MATERIALS AND METHODS

2.1. Database Set. A series of 79 HEPT derivatives, as listed in Table 1, are subjected to QSAR analysis. These compounds, synthesized first by Tanaka et al.[6−9] were studied by several authors[16−18] and discussed in previous studies by us and others.[14,22] Their general structure is presented in Figure 1.

In Table 1, EC$_{50}$ is the effective concentration and refers to the concentration required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (HTLV-III$_B$ strain). Attempts have been made to correlate this activity with a huge number of descriptors encoding the steric, hydrophobic, electronic, and structural features of HEPT derivatives.[16−18] We have gathered those claimed as being relevant in describing the anti-HIV-1 activity variation of the series under investigation. Twenty molecular descriptors have been used as initial data to derive a new QSAR model.

2.2. Neural Networks. NNs are artificial systems emulating the function of the brain where a very high number of information-processing neurons are interconnected. Although there are a number of different NN models,[23] the most frequently used type of NN in QSAR, and the one we shall use in this paper, is the three-layered feed-forward network. In this type of network the neurons are arranged in layers (an input layer, one hidden layer, and an output layer), Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer.

According to the supervised learning adopted in this work, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A back-propagation (BP) algorithm[24] is used to minimize the error function.

3. RESULTS AND DISCUSSION

3.1. Variable Selection. Determination of the relevant variables that tightly describe dependencies between activity and chemical structures of compounds is of a great interest in a QSAR model establishment. More often than not, interrelation between variables skews the results, and, therefore, fortuitous or artful QSAR models may be obtained.

After collecting the data, we submitted all 20 variables to MLR. We included all 79 molecules described in Table 1. The stepwise MLR procedure based on the forward-selection and backward-elimination methods is used to select the powerful variables. To avoid all difficulties in the interpretation of the resulting models, pairs of variables with a correlation coefficient greater than 0.75 were classified as interrelated and only one of these was included in the screened model. A few suitable models were obtained. Furthermore, inspection of an eventual nonlinear correlation between variables has been performed using NN as a nonlinear modeling tool and the resulting correlation coefficients are given in Table 2. This table shows low nonlinear correlation coefficients between the variables ($R^2 < 0.75$). Thus, we end up with a model with seven variables. Table 3 presents the notation and a short description of the molecular descriptors used to generate the QSAR model.

The equation of the best linear fitted model is

$$\log(1/EC_{50}) = -12.059 - 2.188\ B1(R_2) + 0.426\ \log P + 2.137(1/S) - 0.432\ MR(R_2) + 66.279\ {}^4\chi^N_p - 0.868\ B1(3 - R_1) - 23.501\ {}^6\chi^v_{ch}$$

$$N = 79 \quad R^2 = 0.871 \quad S = 0.506 \quad q^2 = 0.840$$

The obtained model points out the main structural features influencing the anti-HIV-1 activity of HEPT derivatives.[22] In addition to the hydrophobic effect (logP), the shape ($^1/_S$) and the branching character of the whole molecule ($^4\chi^N_p$,

Table 2. Nonlinear Correlation Coefficients between the Variables

	$^4\chi^N_p$	B1(3-R_1)	$^6\chi^v_{ch}$	B1(R_2)	logP	MR(R_2)	1/*S*
$^4\chi^N_p$	1.000	0.117	0.626	0.018	0.540	0.595	0.184
B1(3-R_1)	0.034	1.000	0.488	0.017	0.018	0.078	0.132
$^6\chi^v_{ch}$	0.171	0.121	1.000	0.009	0.532	0.086	0.014
B1(R_2)	0.121	0.020	0.014	1.000	0.122	0.043	0.077
logP	0.549	0.019	0.485	0.072	1.000	0.365	0.032
MR(R_2)	0.273	0.065	0.146	0.023	0.159	1.000	0.093
$^1/_S$	0.212	0.171	0.012	0.052	0.022	0.190	1.000

Table 3. Molecular Descriptors Used in the Generation of the QSAR Model

notation	molecular descriptors
logP	octanol/water partition coefficient calculated for the whole molecule
B1(R_2)	STERIMOL descriptor (minimum width of the R_2 substituent)
1/*S*	reciprocal of the standard shadow area on YZ plane
MR(R_2)	molecular refractivity of the R_2 substituent
$^6\chi^v_{ch}$	chain-type molecular connectivity index
$^4\chi^N_p$	path-type molecular connectivity index to the fourth order
B1(3-R_1)	STERIMOL descriptor (minimum width of the R_1 substituent at the 3-position)

$^6\chi^v_{ch}$), there are three main effects: the length and the polarizability of the R_2 substituent (B1(R_2) and MR(R_2)), and the length of the R_1 substituent in position 3 (B1 (3-R_1)).

In the following step, we used these data for building a three-layered NN employing BP learning strategy.

3.2. Neural Network Model. *3.2.1. Training Stage.* The architecture of a network is the main feature influencing the flexibility of the model it generates.[15] In our case a three-layer NN has been used. Seven neurons constitute the input layer and describe the seven variables chosen by the MLR analysis and strengthened by NN. One neuron, which encodes the anti-HIV-1 activity, constitutes the output layer and the hidden layer contains a variable number of neurons. It has been suggested that a ratio, ρ, plays a crucial role in determining the number of hidden units being employed.[24] ρ is defined as

$$\rho = \frac{\text{number of examples presented to the network}}{\text{number of connections in the network}}$$

The number of hidden neurons was varied from three to eight in order to maintain the ρ ratio in the [1.0−3.0] range. It is claimed that, for $\rho \ll 1.0$, the network simply memorizes the data. For $\rho \gg 3.0$, the network is not able to generalize.[25] A bias term was added to the input and the hidden layers. The input values were normalized to [0.1−0.9] interval. The sigmoid function was used as the transformation function, and the delta rule[26] was used as the error correction formula. The weights were initialized to random values between −0.5 and +0.5 and no momentum was added. The learning rate was initially set to 1 and was gradually decreased during training. A BP algorithm implemented in C language was developed.

Six NN architectures of 7-*x*-1 ($x = 3-8$) were then trained. The optimal number of iterations required was 10 000. The results of QSAR done by these NN architectures and by MLR analysis are listed in Table 4. The quality of the fitting is estimated by the standard error of calculation (*S*) and by the correlation coefficient (R^2).

From Table 4 one can easily notice that all NN architectures trained showed high fitting ability. The high correlation coefficients given by the trained NN architectures indicate that the log(1/EC_{50}) activity was significantly correlated with the seven variables adopted in this work.

Table 4. Statistical Results of Different NN Architectures; Results of MLR Analysis Are Also Given

architecture	R^2	*S*
7-3-1	0.962	0.260
7-4-1	0.973	0.220
7-5-1	0.977	0.203
7-6-1	0.986	0.159
7-7-1	0.988	0.145
7-8-1	0.993	0.111
MLR	0.871	0.506

It is noteworthy that the results of the NN are significantly better than those obtained by MLR analysis. Usually, nonlinear QSAR problems show larger improvements when modeled by NN.[27] This provides evidence for the nonlinearity of the relationship between structural features of HEPT compounds and anti-HIV-1 activity.

3.2.2. Prediction Stage. It is obvious that medicinal chemists take a keen interest in the design of new drugs. What is needed is a system that is able to provide reasonable predictions for the compounds that are previously unknown. Indeed, one of the most important attributes of NNs is their ability to generalize,[28] that is their ability to make predictions on new data with accuracy similar to that with the training set. Besides, NNs are known for their ability to model a wide set of functions without knowing the analytic forms in advance. After training, the NN is initiated to recognize the relationship between input and output data and creates an internal model as a governing data process. The NN can then use this internal model to make predictions for new inputs.

In addition to the estimation of predictive ability of NN, a leave-one-out (LOO) procedure[29] was used to estimate which NN architecture would be best suited for the data set of the 79 HEPT derivatives or which NN architecture would result in a QSAR model having the lowest prediction error. The analysis of predictive ability was carried out in terms of both predictive r^2 and standard error of prediction S_{press}.

Predictive r^2 is defined as

$$r^2 = \frac{\text{SSY} - \text{PRESS}}{\text{SSY}}$$

where SSY is the sum of squares of deviations of the

Table 5. Predictive Ability of NN; Results of MLR Analysis Are Also Given

architecture	predictive r^2	S_{press}
7-3-1	0.858	0.503
7-4-1	0.850	0.517
7-5-1	0.862	0.497
7-6-1	0.851	0.518
7-7-1	0.855	0.508
7-8-1	0.846	0.523
MLR	0.840	0.533

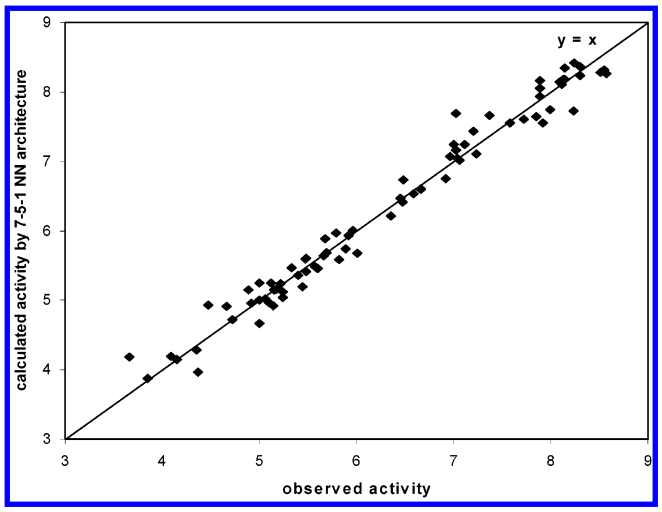

Figure 2. log(1/EC$_{50}$) Calculated by 7-5-1 NN architecture versus log(1/EC$_{50}$) observed experimentally.

observed values from their mean and PRESS is the prediction error sum of squares, derived from the LOO procedure. S_{press} is defined as

$$S_{press} = \sqrt{\frac{PRESS}{N}}$$

where N is the number of compounds used in this study. The maximum number of iterations was set to 1000, but sufficient convergence was usually observed, at least for small networks, after 200 iterations.

As shown in Table 5 all NN architectures submitted to the LOO procedure gave good results, yielding high predictive r^2 values.

Table 4 shows that the 7-8-1 NN architecture (containing 73 connections) gave the lowest standard error of calculation in the training stage. However, its ability to generalize (Table 5) gets bad. On the other hand, the 7-5-1 NN architecture, with 46 connections, exhibited a good predictive performance. Besides, its fitting ability seems satisfying. The ability to generalize being the most important factor, we have come to the conclusion that the 7-5-1 architecture is well adapted to relate anti-HIV-1 activity of HEPT derivatives to their structural requirements, so it was adopted for the analyses that will follow.

The values of the log(1/EC$_{50}$) calculated by 7-5-1 NN architecture are plotted against the log(1/EC$_{50}$) observed experimentally, and are illustrated in Figure 2. This figure shows that the log(1/EC$_{50}$) values calculated by the NN are very close to the experimental ones.

To take a closer look at the resulting NN model, a residual analysis was performed (Table 1). A normal residual plot

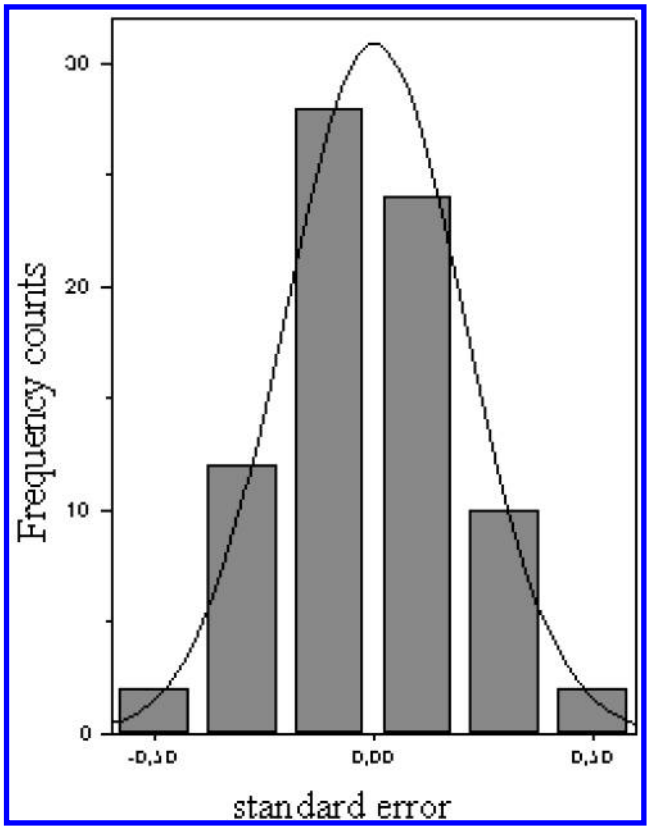

Figure 3. Normal residual plot obtained from the 7-5-1 architecture.

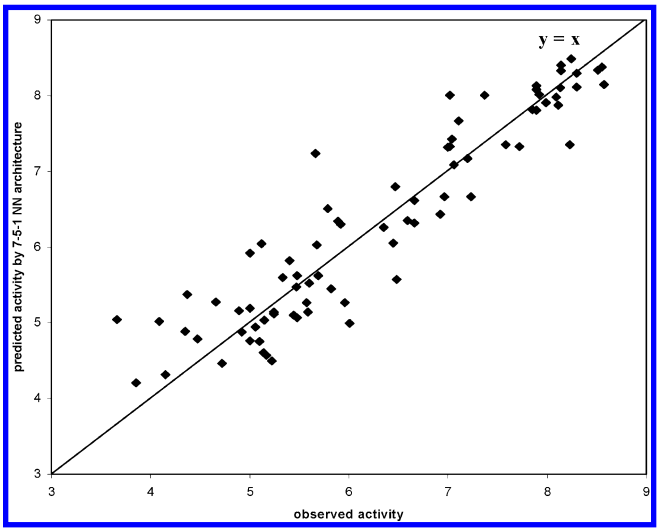

Figure 4. Predicted versus experimental log(1/EC$_{50}$) values.

was obtained and it is depicted in Figure 3. An examination of this figure shows that the outputs of NN coincide with the target outputs to a higher degree. The NN's calculated values are correct for a large number of molecules.

The predictive power of 7-5-1 NN architecture can be judged from the plot of predicted versus experimental log(1/EC$_{50}$) values shown in Figure 4.

All through the analyses above, we have related anti-HIV-1 activity of HEPT derivatives to seven descriptors referring to their main structural features. Table 6 shows the statistical results obtained in this work compared to those obtained in other works.[14,16−18,21] We can assume that the QSAR model

Table 6. Comparison Between Some Works on the Same Set of HEPT Derivatives

reference	model	R^2	S	predictive r^2	number of descriptors
16	MLR	0.900	0.439	0.745	9
	PLS	0.889	0.440	0.8599	9
17	MLR	0.815	0.600	0.783	5
18	MLR	0.811	0.607	0.778	6
	NN	0.919	0.270		6
21	MLR	0.830	0.506	0.700	4
	NN	0.852	0.458	0.810	4
14	NN	0.984	0.165	0.853	8
current work	NN	0.977	0.203	0.862	7

derived here outperforms, in terms of predictive ability, those given in the literature. It is worth mentioning that there is a slight improvement in the predictive performance of the NN model (constructed with 7 molecular descriptors) with regard to the PLS model (constructed with 9 molecular descriptors). The seven descriptors fed the 7-5-1 NN with the information necessary to describe the structural variations of HEPT derivatives. In addition, the present model takes the advantage in that the nonlinear aspect of the relationship between the variables and the activity is well handled by NN, as can be seen in the next section.

3.2.3. Effect of the Molecular Properties on the Anti-HIV-1 Activity. One of the purposes of QSAR analyses is to understand the forces governing the activity of a particular class of compounds and to assist drug design. Therefore, the evaluation of the relevance of the descriptors proved quite interesting and useful to shed more light on the structure−anti-HIV-1 activity. That is why we chose to estimate their relative contribution. The contribution of each descriptor to the establishment of the QSAR has been estimated from the trained 7-5-1 NN. We excluded descriptor i together with its corresponding weights from the 7-5-1 NN and retrained the resulting 6-5-1 NN as usual. The mean of the deviations' absolute values Δe_i between the experimental anti-HIV-1 activity and the estimated activity for all compounds has been calculated. This process has been reiterated for each descriptor. Finally, the contribution (C_i) of the descriptor i is given by

$$C_i = \frac{100 \cdot \Delta e_i}{\sum_{i=1}^{7} \Delta e_i}$$

Figure 5 shows the contribution of each descriptor.

According to the plot of Figure 5, the descriptors related to both hydrophobic and steric characters seem to be very relevant in the establishment of the quantitative structure−anti-HIV-1 activity relationship of HEPT derivatives. There are two main reasons that could explain the important contribution of these descriptors compared to the others: (i) the role of the hydrophobicity of these compounds in the inhibition mechanism of the RT; and (ii) the role of the shape of molecules in their interaction with the RT enzyme.

These hypotheses are supported by numerous reports[3,4] stating that the pocket in which all NNRTIs bind is predominantly hydrophobic with substantial aromatic character (Tyr 181, Tyr 188, Phe 227, Trp 229, and Tyr 232). Moreover, structural information has been obtained for nonnucleoside inhibitors complexed with RT,[30] as well as for unliganded RT.[4] These structures have provided a great deal of insight into the conformational flexibility of RT, including the structural changes that are induced by binding of NNRTIs. These results strongly reveal the importance of the steric feature of the inhibitors in their interaction with the binding pocket. Interestingly, the present model performed by NN reproduces well this behavior. We can notice the variation of anti-HIV-1 activity of molecules according to the $^1/_S$ descriptor from the plot of Figure 6.

The plot of Figure 6 has been obtained using a technique proposed by So and Richards[25] and Andrea and Kalayeh.[24] We monitored the variation of the network output by changing the value of the $^1/_S$ input, while keeping the other

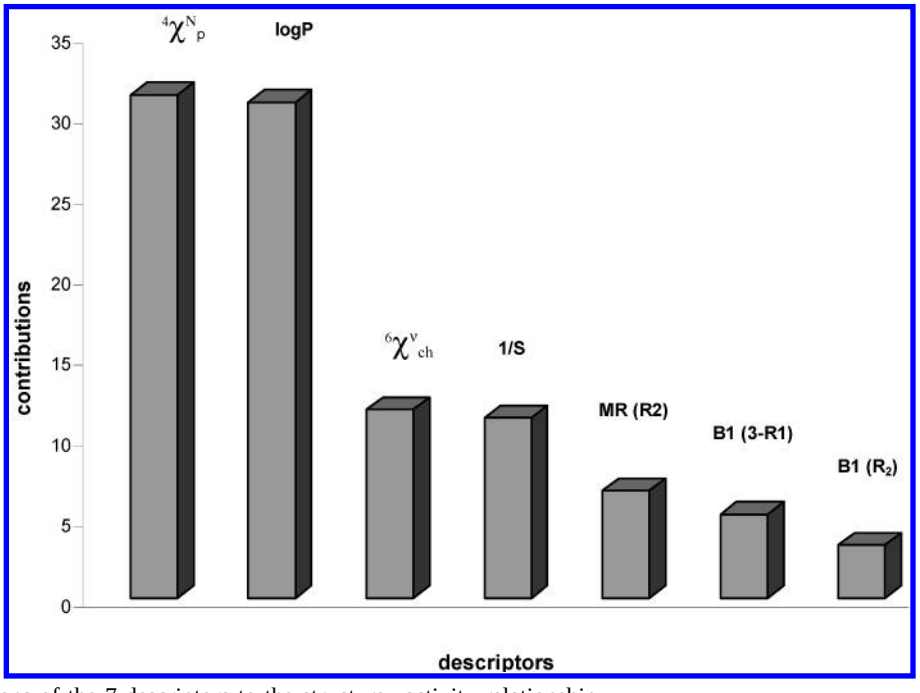

Figure 5. Contributions of the 7 descriptors to the structure−activity relationship.

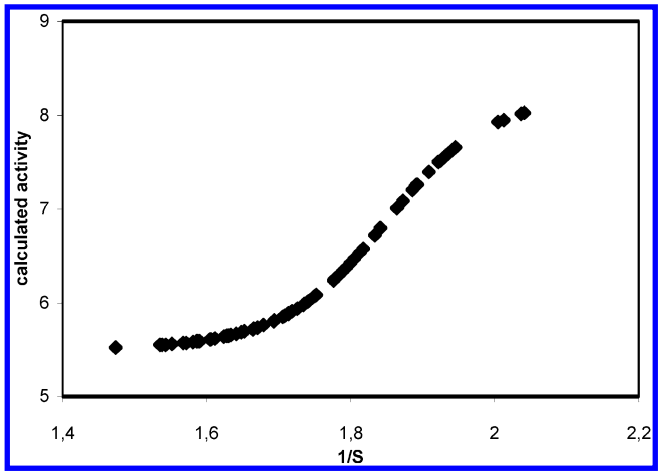

Figure 6. Calculated anti-HIV-1 activity versus $^1/_S$.

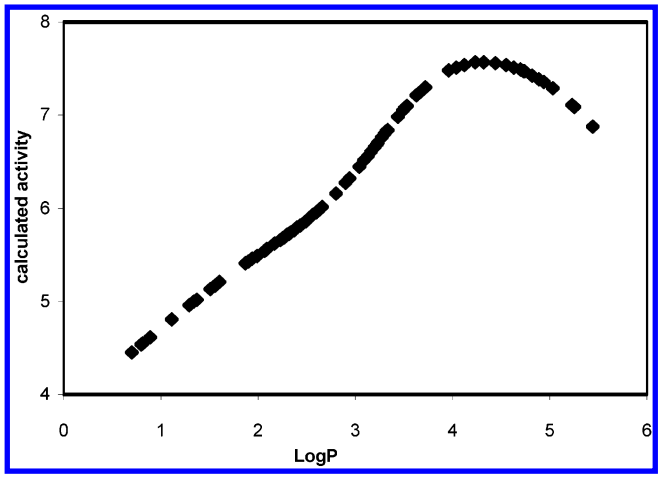

Figure 7. Calculated anti-HIV-1 activity versus logP.

inputs constant. From this analysis we can infer some structural requirements of the RT binding pocket. Whereas moderately sized molecules induce favorable steric interactions with the binding pocket, there is not enough space to accommodate molecules which are too large, and smaller ones has sufficient steric requirements to interact with amino acid residues.

The same procedure has been used for the hydrophobic logP descriptor. The resulting plot is depicted in Figure 7.

From this figure, a parabolic dependence of anti-HIV-1 activity to the logP can be easily noticed. The activity increases with hydrophobicity until the hydrophobic optimum of 4.32 is reached and then it descends. Note that the optimum hydrophobic of 4.32 lies in the [2.31−5.72] range determined by Garg et al.[17] for the class of NNRTIs. Moreover, this value is not far away from 4.95 of Efavirenz (Sustiva),[17] already approved by the FDA. This figure stresses the important role that hydrophobicity plays not only in maximizing ligand−receptor interactions but also in drug transport and permeability phenomena. One can think that the increase of the hydrophobicity of the synthesized compounds can lead to increase of the anti-HIV-1 activity. However, this increase can disturb the transport of the compounds through the cellular membranes.

4. CONCLUSION

In this paper, we have applied an NN to the study of a nonlinear QSAR for HEPT derivatives acting as NNRTIs. It was demonstrated that a fully interconnected three-layered NN trained with the BP procedure could learn the correct association between seven relevant descriptors of HEPT derivatives and $\log(1/EC_{50})$. Once it has been trained, the NN with five hidden neurons exhibited a high predictive power. It succeeded in predicting the anti-HIV-1 activity of HEPT derivatives which were not included in the model development set. The computed anti-HIV-1 activity values are correct for a large number of studied compounds. The descriptors used describe well the structural requirements of HEPT derivatives. The contributions of the descriptors to the QSAR were evaluated. They confirmed the well-known role of hydrophobic and steric effects in the interaction between NNRTIs and their target pocket in the RT enzyme. The nonlinear relationship between these factors and the anti-HIV-1 activity has been clearly shown. The results show that the model is a reliable one and it may be of considerable interest for the design of new anti-HIV-1 drugs.

Note Added after ASAP Posting. This article was released ASAP on 6/12/2003 with $\log P(1/EC_{50})$ instead of $\log(1/EC_{50})$ in the caption of Figure 4 and text. In addition, in Table 6, third column line 6 appeared as 830 instead of 0.830. The correct version was posted on 6/24/2003.

REFERENCES AND NOTES

(1) Turner, B. G.; Summers, M. F. Structural Biology of HIV. *J. Mol. Biol*. **1999**, *285*, 1−32.

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

(3) De Clercq, E. Perspective of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) in the Therapy of HIV-1 Infection. *Il Farmaco* **1999**, *54*, 26−45.

(4) Esnouf, R.; Ren, J.; Ross, C.; Jones, Y.; Stammers, D.; Stuart, D. Mechanism of Inhibition of HIV-1 Reverse Transcriptase by Non-Nucleoside Inhibitors. *Struct. Biol.* **1995**, *2*, 303−308.

(5) Miyasaka, T.; Tanaka, H.; Baba, M.; Hayakawa, H.; Walker, R. T.; Balzarini, J.; DeClercq, E. A Novel Lead for Specific anti-HIV-1 Agents: 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1989**, *32*, 2507−2509.

(6) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Balzarini, J.; DeClercq, E. A New Class of HIV-1 Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV Activity of 5- or 6- Substituted Analogues of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349−357.

(7) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Miyasaka, T. Synthesis and Anti-HIV Activity of 2-,3-, and 4-Substituted Analogues of 1-[(2-hydroxy-ethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem*. **1991**, *34*, 1394−1399.

(8) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; DeClercq, 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 the C-5 Position on Anti-HIV-1 Activity. *J. Med. Chem.* **1992**, *35*, 337−345.

(9) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; DeClercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of Deoxy Analogues of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT) as Potent and Selective Anti-HIV Agents. *J. Med. Chem*. **1992**, *35*, 4713−4719.

(10) Hansh, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*. American Chemical Society: Washington, DC, 1995.

(11) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem*. **1990**, *33*, 2583−2590.

(12) Zhu, J.; Chen, D.; Wu, P. Exploration of Artificial Neural System for Simulation of the Halogen-Exchange Fluorination Reaction. *Comput. Chem*. **1999**, *23*, 97−100.

(13) Yan, A.; Jiao, G.; Hu, Z.; Fan, B. T. Use of Artificial Neural Networks to Predict the Gas Chromatographic Retention Index Data of Alkylbenzenes on Carbowax-20M. *Comput. Chem.* **2000**, *24*, 171−179.
(14) Douali, L.; Villemin, D.; Cherqaoui, D. Comparative QSAR based on Neural Networks for the Anti-HIV Activity of HEPT Derivatives. *Curr. Pharm. Des.*, in press.
(15) Zupan, J.; Novic, M. General Type of a Uniform and Reversible Representation of Chemical Structures. *Anal. Chim. Acta* **1997**, *348*, 409−418.
(16) 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.
(17) Garg, R.; Gupta, S. P.; Gao, H.; Babu, M. S. Debnath, A. K. Comparative Quantitative Structure−Activity Relationship Studies on Anti-HIV Drugs. *Chem. Rev.* **1999**, *99*, 3525−3601.
(18) Jalali-Heravi, M.; Parastar, F. Use of Artificial Neural Networks in a QSAR Study of Anti-HIV Activity for a large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147−154.
(19) Kireev, D. B.; Chretien, J. R.; Griersson, D. S.; Monneret, C. A. 3D QSAR Study of a Series of HEPT Analogues: The Influence of Conformational Mobility on HIV-1 Reverse Transcriptase Inhibition. *J. Med. Chem.* **1997**, *40*, 4257−4264.
(20) Hannongbua, S.; Nivesnanond, K.; Lawtrakul, L.; Pungpo, P.; Wolschann, P. 3D-Quantitative Structure−Activity Relationships of HEPT Derivatives as HIV-1 Reverse Transcriptase Inhibitors, Based on Ab Initio Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 848−855.
(21) Bazoui, H.; Zahouily, M.; Boulajaaj, S.; Sebti, S.; Zakarya, D. QSAR for anti-HIV activity of HEPT derivatives. *SAR QSAR Environ. Res.* **2002**, *13*, 567−577.
(22) Gaudio, A. C.; Montanari, C. A. HEPT Derivatives as Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase: QSAR Studies agree with the Crystal Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 287−295.
(23) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists.* VCH−Wiley: Weinheim, Germany, 1993; p 207.
(24) Andrea, T. A.; Kalayeh, H. Application of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(25) So, S.; Richards, G. Application of Neural Networks: Quantitative Structure−Activity Relationships of the Derivatives of 2,4-Diamino-5- (substituted-benzyl) pyrimidines as DHFR Inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.
(26) Freeman, J. A.; Skapura, D. M. *Neural Networks Algorithms, Applications and Programming Techniques*, Addison Wesley Publishing Company: Reading, MA, 1991.
(27) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.
(28) Bishop, C. M. Neural Networks and their Applications. *Rev. Sci. Instrum.* **1994**, *65*, 1803−1832.
(29) Zakarya, D.; Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cense, J. M. Application of Neural Networks to Structure-Sandalwood Odour Relationships. *J. Phys. Org. Chem.* **1997**, *10*, 612−622.
(30) Ren, J.; Esnouf, R.; Hopkins, A.; Ross, C.; Jones, Y.; Stammers, D.; Stuart, D. The Structure of HIV-1 Reverse Transcriptase Complexed with 9-Chloro-TIBO: Lessons for Inhibitor Design. *Structure* **1995**, *3*, 915−926.

CI034047Q