Development of Neural Network QSPR Models for Hansch Substituent Constants. 2. Applications in QSAR Studies of HIV-1 Reverse Transcriptase and Dihydrofolate Reductase Inhibitors

Ting-Lan Chiu and Sung-Sau So*

Roche Research Center, Hoffmann-La Roche Inc., Nutley, New Jersey 07110

Received June 24, 2003

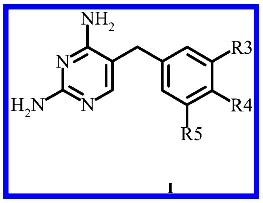
In this paper, the applications of a Hansch substituent constant predictor[1] to Quantitative Structure−Activity Relationships (QSAR) studies of *E. coli* dihydrofolate reductase (DHFR) inhibitors 2,4-diamino-5-(substituted-benzyl)pyrimidines as well as HIV-1 reverse transcriptase (RT) inhibitors 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives are demonstrated. Both data sets contain functional groups for which the substituent constants (π, MR, F and R) could not be found in standard substituent constant tables. The substituent constant predictor allowed us to derive predicted π, MR, F and R values for all substituents in both data sets, thus enabling the generation of easily interpretable QSAR models of comparable or better predictivity than previous models.

INTRODUCTION

In our previous paper, a prediction system based on neural network Quantitative Structure−Property Relationship (QSPR) models for substituent constants π, MR, F and R was developed.[1] This type of descriptor is very useful for Quantitative Structure Activities Relationship (QSAR) studies because discovery chemists are familiar with the concept of hydrophobicity, steric hindrance, and electronic effects in molecular recognition. A method of predicting these constants is desirable because it is quite often the case that the constants for some less common substituents in a given chemical series have not been tabulated, making it impossible to use such descriptors to construct a QSAR model for activity correlation.[1]

To facilitate the use of this new prediction tool, a Web interface similar to that described by Ertl[2,3] has been created. In our implementation, the substituent constant predictor Web-page returns the calculated values of π, MR, F and R from neural network QSPR models when a list of substituents expressed in SMILES notation is submitted. In this paper, two QSAR applications using the calculated substituent constants are demonstrated in the following inhibitory activity data sets: *E. coli* dihydrofolate reductase (DHFR) inhibitors 2,4-diamino-5-(substituted-benzyl)pyrimidines as well as HIV-1 reverse transcriptase (RT) inhibitors 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives. These two data sets were chosen because they have been extensively studied; moreover, the substituent constants of some functional groups in these two data sets are not available. It is the goal of this paper to show that, using the substituent constant predictor, new substituent constants are derived and furthermore, robust QSAR models of comparable or better predictivity than other published models can be obtained.

1. DHFR Inhibitors 2,4-Diamino-5-(substituted-benzyl)-pyrimidine Derivatives. DHFR inhibitors 2,4-diamino-5-(substituted-benzyl)pyrimidine derivatives (I) differ from each other at the R_3, R_4, and R_5-positions.



I

To obtain QSAR models with significant correlations, Hansch and co-workers suggested the use of truncated MR (denoted as MR′) as an independent variable in their QSAR analysis of DHFR inhibitors 2,4-diamino-5-(substituted-benzyl)pyrimidine derivatives.[4] In their work, the upper limit of MR is 0.79. The best QSAR model constructed from 68 compounds is a multiple linear regression (MLR) model based on the four descriptors: π_3, MR$_3'$, MR$_4$, MR$_5'$.[5] The correlation coefficient, r, is 0.89. So and Richards applied neural networks to the same data set and improved the Spearman rank correlation coefficient (SRCC) from 0.88 (regression analysis) to 0.94.[6] Hirst used a nonlinear QSAR correlation method and obtained comparable results to those with MLR or ANN.[7] Both So et al. and Hirst constructed their QSAR models based on the same four descriptors. The results of other QSAR studies[8−15] are summarized in Table 1.

The use of truncated MR values with the addition of only the size and hydrophobic parameters considered as descriptors in these QSAR studies highlights two issues. First, the choice of functional transformation for MR values, MR′ = min(MR,0.79), was the result of the application of expert knowledge. A logical question then is what the results would be if no expert knowledge were involved and absolute MR values were used. Second, the four-descriptor model, {π_3,

* Corresponding author phone: (973)235-2193; fax: (973)235-2682; e-mail: sung-sau.so@roche.com.

10.1021/ci030294i CCC: $27.50 © 2004 American Chemical Society
Published on Web 12/16/2003

Table 1. A Summary of Previous QSAR Studies on DHFR Inhibitors Pyrimidines

author	data set size	descriptors	correlation method	major results	ref
Dietrich et al.	23	substituent constant	MLR	$r^2 = 0.84$	4
Li et al.	36	substituent constant	MLR	$r^2 = 0.52$	8
Hansch et al.	44	substituent constant	MLR	$r^2 = 0.65$	9
Selassie et al.	68	substituent constant	MLR	$r^2 = 0.79$	5
So et al.	68 (49/19)[a]	substituent constant	ANN	SRCC = 0.94, $r_{cv}^2 = 0.724$	6
Hirst et al.	74(55/19)[a]	physicochemical	GOLEM	SRCC for training: cv: test = 0.948:0.692:0.738	10
Hirst	74(55/19)[a]	substituent constant	nonlinear QSAR	SRCC for training: cv: test = 0.80:0.68:0.59	7
King et al.	55	substituent constant	stepwise linear regression	$r_{cv}^2 = 0.84$	11
Loukas	68(48/10/10)[b]	whole-molecule; atom and substituent	ANFIS	q^2(test) = 0.909	12
Landavazo et al.	55	physicochemical	evolved neural networks	SRCC = 0.724	13
Hasegawa et al.	11	CoMFA	quadratic PLS	q^2 (leave-one-out cv) = 0.940	14
Seri-Levy et al.	26	shape similarity	linear	$r^2 = 0.926$	15

[a] The first and second numbers in parentheses are the numbers of compounds in the training and test sets, respectively. [b] The three numbers separated by slashes denote the numbers of compounds in the training, validation and test sets, respectively.

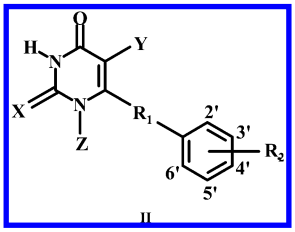
Table 2. A Summary of Previous QSAR Studies on HEPT Derivatives

data set size	descriptor type(s)	correlation method	major results	ref
33	Hansch	MLR	$r^2 = 0.89$	15
36	Hansch; indicator	MLR	$r^2 = 0.90$, $r_{cv}^2 = 0.84$	17
35	Hansch; indicator; structural	MLR	$r^2 = 0.84$, $r_{cv}^2 = 0.79$	17
87	3D	MLR	$r^2 = 0.88$	18
107	topological, structural; Hansch; indicator	MLR	$r^2 = 0.90$, $r_{cv}^2 = 0.67$	20
101	3D	PLS	$r^2 = 0.93$, $r_{cv}^2 = 0.84$	19
95	physicochemical	MLR	$r^2 = 0.83$, $r_{cv}^2 = 0.70$	21
95	physicochemical	ANN	$r^2 = 0.85$, $r_{cv}^2 = 0.81$	21

MR_3', MR_4, MR_5'}, was the best model when only hydrophobic and size parameters were considered and truncated MR values were used. It is also of interest to know whether the best model would be the same if additional electronic parameters, such as F and R, became available and absolute MR values were used.

In this paper, we attempt to address these issues by use of our Hansch substituent constant predictor. Specifically, the predictor enables us to obtain π, MR, F and R values for all substituents at the R_3, R_4 and R_5 positions of 2,4-diamino-5-(substituted-benzyl)pyrimidines in the data set. The results of our studies show that comparable QSAR models can be derived without preprocessing of descriptors, i.e., no expert knowledge is involved. In addition, we demonstrate that electronic contributions also seem to play an important role in this system, as reflected by the choice of descriptors in the best four-descriptor model from our analysis.

2. HIV-1 Reverse Transcriptase Inhibitors 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) Derivatives. HIV-1 reverse transcriptase (RT) inhibitor HEPT derivatives (II) differ from each other at the X, Y, Z, R_1 and R_2 positions. Specifically, the 2′-6′ positions of R_2 can be substituted or unsubstituted.



II

Recently, a review article on QSAR studies of the HEPT series has been published, summarizing various aspects of QSAR models and their consistency with the X-ray structures of HIV-1 RT − HEPT complexes.[16] The principal conclusions of the previous QSAR studies are as follows. First, the substituents at the Z position are preferably limited in size and not branched. Phenyl is favorable because phenyl stacking interactions with nearby aromatic residues seem to play a role. Second, the substituents at the Y position are preferably limited in size, branched, and lipophilic. Isopropyl is favorable. Furthermore, the 2′, 4′ and 6′ positions of the R_2 ring are preferably unsubstituted, while the 3′ and 5′ positions of the ring are preferably substituted.[17−22] The best two QSAR models constructed using 101 or more HEPT analogues demonstrated an r^2/q^2 of 0.90/0.67[22] and 0.93/0.84,[21] respectively. A summary of previous QSAR studies, including very recent work done by Bazoui et al.,[23] can be found in Table 2.

METHODS

1. Data Sets. *DHFR Inhibitors: Pyrimidine Derivatives.* The data set was taken from ref 5 and has been extensively studied by previous QSAR studies,[5,6,11,12] thus providing us with an opportunity to compare our QSAR model to previously derived models. The substituent $3,4{-}OCH_2O{-}$ was removed from our data set because it forms a fused ring with the template. As mentioned in the model development section of the companion paper, the current version of our Hansch substituent predictor does not handle linker-type groups with multiple attachment points. The final data set, consisting of 67 pyrimidine derivatives, is shown in Table 3.

HIV-1 RT Inhibitors: HEPT Derivatives. The data set was compiled from two references[19,22] and is shown in Table 4. This data set has also been extensively studied by previous

Table 3. Data Set of Pyrimidine Derivatives

no.	R_3	R_4	R_5	pK_i
1	CH_2CH_3	CH_2CH_3	CH_2CH_3	7.82
2	OCH_3	$OCH_2CH_2OCH_3$	OCH_3	8.35
3	OCH_3	OCH_3	OCH_3	8.08
4	OCH_3	$N(CH_3)_2$	OCH_3	7.71
5	OCH_3	Br	OCH_3	8.18
6	OCH_3	SCH_3	OCH_3	8.07
7	OCH_3	$C(CH_3){=}CH_2$	OCH_3	8.12
8	OCH_2CH_3	1H-pyrrol-1-yl	OCH_2CH_3	7.66
9	OCH_3	$O(CH_2)_7CH_3$	OCH_3	7.20
10	CH_2OH	H	CH_2OH	6.31
11	OCH_3	H	OCH_3	7.71
12	OCH_2CH_3	H	OCH_2CH_3	7.69
13	OCH_2CH_3	H	$OCH_2CH_2CH_3$	7.69
14	$OCH_2CH_2CH_3$	H	$OCH_2CH_2CH_3$	7.41
15	CH_3	H	CH_3	7.04
16	H	OH	OH	6.46
17	H	$NHCOCH_3$	NO_2	6.97
18	H	$OCH_2CH_2OCH_3$	$OCH_2CH_2OCH_3$	7.22
19	H	OCH_3	OCH_3	7.72
20	H	OCH_3	OH	6.84
21	H	OSO_2CH_3	OCH_3	7.94
22	H	OH	OCH_3	7.54
23	H	$OCH_2CH_2OCH_3$	OCH_3	7.77
24	H	$OCH_2C_6H_5$	OCH_3	7.53
25	H	OCH_3	OSO_2CH_3	7.80
26	$OCH_2C_6H_5$	OCH_3	H	7.66
27	CF_3	OCH_3	H	7.69
28	$O(CH_2)_7CH_3$	OCH_3	H	7.16
29	OCH_2CH_3	$OCH_2C_6H_5$	H	7.35
30	H	H	OCH_2CONH_2	6.57
31	H	H	CH_2OH	6.28
32	H	H	OSO_2CH_3	6.92
33	H	H	CH_2OCH_3	6.59
34	H	H	OH	6.47
35	H	H	$OCH_2CH_2OCH_3$	6.53
36	H	H	OCH_3	6.93
37	F	H	H	6.23
38	CH_3	H	H	6.70
39	Cl	H	H	6.65
40	Br	H	H	6.96
41	CF_3	H	H	7.02
42	$CO(CH_2)_3CH_3$	H	H	6.55
43	I	H	H	7.23
44	$O(CH_2)_3CH_3$	H	H	6.82
45	$OCH_2C_6H_5$	H	H	6.99
46	$O(CH_2)_5CH_3$	H	H	6.86
47	$O(CH_2)_6CH_3$	H	H	6.39
48	$O(CH_2)_7CH_3$	H	H	6.25
49	H	NH_2	H	6.30
50	H	$NHCOCH_3$	H	6.89
51	H	OSO_2CH_3	H	6.60
52	H	OH	H	6.45
53	H	$OCH_2CH_2OCH_3$	H	6.40
54	H	NO_2	H	6.20
55	H	OCH_3	H	6.82
56	H	F	H	6.35
57	H	$N(CH_3)_2$	H	6.78
58	H	CH_3	H	6.48
59	H	Cl	H	6.45
60	H	Br	H	6.82
61	H	OCF_3	H	6.57
62	H	$O(CH_2)_3CH_3$	H	6.89
63	H	$OCH_2C_6H_5$	H	6.89
64	H	$O(CH_2)_5CH_3$	H	6.07
65	H	$O(CH_2)_6CH_3$	H	6.10
66	H	C_6H_5	H	6.93
67	H	H	H	6.18

QSAR studies, providing us with a further opportunity to compare our QSAR models with other published results. The full data set consists of a total of 93 compounds that was further divided into a training set of 75 compounds (1−75 in Table 4) and a test set of 18 compounds (76−93 in Table 4).

2. Descriptor Generation. *DHFR Inhibitors: Pyrimidine Derivatives.* The calculated sets of π, MR, F, and R values of all substituents at the R_3, R_4, and R_5 positions were obtained by using our Hansch substituent predictor Web interface. As mentioned in the Introduction, the calculated values were derived from neural network QSPR models. As a result, twelve calculated descriptors, $\{\pi_3, MR_3, F_3, R_3, \pi_4, MR_4, F_4, R_4, \pi_5, MR_5, F_5, R_5\}$, were generated for activity correlation in this QSAR study.

HIV-1 RT Inhibitors: HEPT Derivatives. A few of the crucial descriptors used in previous QSAR studies were adopted and combined with the Hansch substituent constants for the Z, Y and R_2 positions, to generate a pool of 18 descriptors for each derivative.[19,22] A summary of these 18 descriptors is given in Table 5. Similar to the DHFR study, the calculated π, MR, F, and R values of all substituents for the R_2, Y and Z positions were obtained by using our Hansch substituent predictor Web interface.

3. Correlation Methods. *Linear Regression Methods.* Multiple Linear Regression (MLR) and Partial Least Squares (PLS) were performed using the SIMCA-P+ 10.0 statistical package available from Umetrics AB, Box 7960, S-907 19 Umeå, Sweden.

Nonlinear Method. Artificial Neural Network (ANN) simulations were performed using an in-house package.[24,25] Due to the stochastic nature of ANN simulations, each calculation was the combined result from an ensemble of ten ANN runs using different random initial seeds. Regarding the DHFR inhibitor data set, the number of input variables was fixed to 4 and an ANN of 4-4-1 architecture was established throughout the entire study. For HEPT derivatives, however, the number of the descriptors (input nodes) was varied from 2 to 16. The number of adjustable weights is equal to $H\times(I+O)+H+O$, where H, I and O represent the number of hidden, input and output nodes, respectively. The number of hidden nodes was varied from 7 (for 2 input nodes) to 1 (for 16 input nodes) to ensure that the number of adjustable weights is approximately the same for the networks containing different numbers of input nodes (descriptors). Furthermore, as suggested by Andrea et al.,[26] we made sure that the ratio ρ, defined as the ratio of the number of data points in the training set over the number of adjustable weights, fell in the range of 1.8 to 2.2 in order to avoid over-fitting of data.

4. Evaluation of Correlation. The correlation between the predicted and experimental data is measured by the Pearson correlation coefficient r^2 for the training set, Pearson correlation coefficient r^2 for the test set (if a test set is available), and cross-validated correlation coefficient q^2 for the leave-one-out runs. The definitions of r^2 and q^2 can be found in ref 27.

RESULTS AND DISCUSSION

1. QSAR Studies on Pyrimidine Derivatives. As mentioned in the Introduction, the following two issues were addressed. The first issue is what the results would be if no application of expert knowledge were involved and absolute MR values were used to build the model. The second issue is whether the best model would be the same if additional

Table 4. HEPT Derivatives Data Set[a]

no.	R1	R2	X	Y	Z	$\log(1/EC_{50})$
1	S	2-NO_2	O	Me	$CH_2OCH_2CH_2OH$	3.85
2	S	2-Me	O	Me	$CH_2OCH_2CH_2OH$	4.15
3	S	3-CF_3	O	Me	$CH_2OCH_2CH_2OH$	4.35
4	S	H	O	CH_2Ph	$CH_2OCH_2CH_2OH$	4.37
5	S	3-NO_2	O	Me	$CH_2OCH_2CH_2OH$	4.47
6	CH_2	H	O	Me	$CH_2OCH_2CH_2OH$	4.64
7	S	3-OMe	O	Me	$CH_2OCH_2CH_2OH$	4.66
8	S	2-OMe	O	Me	$CH_2OCH_2CH_2OH$	4.72
9	S	3-Cl	O	Me	$CH_2OCH_2CH_2OH$	4.89
10	S	3-CN	O	Me	$CH_2OCH_2CH_2OH$	5.00
11	S	H	S	Pr	$CH_2OCH_2CH_2OH$	5.00
12	S	H	O	Me	$CH_2OCH_2CH_2OMe$	5.06
13	S	3-COOMe	O	Me	$CH_2OCH_2CH_2OH$	5.10
14	S	H	O	Me	$CH_2OCH_2CH_2OCOPh$	5.12
15	S	3-COMe	O	Me	$CH_2OCH_2CH_2OH$	5.14
16	S	H	O	Me	$CH_2OCH_2CH_2OH$	5.15
17	S	H	O	Me	$CH_2OCH_2CH_2OCOMe$	5.17
18	S	H	O	Me	$CH_2OCH_2CH_2N_3$	5.24
19	S	3-Br	O	Me	$CH_2OCH_2CH_2OH$	5.24
20	S	H	O	Me	CH_2OBu	5.33
21	S	H	O	I	$CH_2OCH_2CH_2OH$	5.44
22	S	H	O	Me	CH_2OPr	5.44
23	S	H	O	Pr	$CH_2OCH_2CH_2OH$	5.47
24	S	3-F	O	Me	$CH_2OCH_2CH_2OH$	5.48
25	S	3-Et	O	Me	$CH_2OCH_2CH_2OH$	5.57
26	S	3-Me	O	Me	$CH_2OCH_2CH_2OH$	5.59
27	S	H	O	Me	Et	5.66
28	S	H	O	Me	CH_2OMe	5.68
29	S	H	O	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	5.69
30	S	H	O	Me	$CH_2OCH_2CH_2Cl$	5.82
31	S	3,5-Cl	O	Me	$CH_2OCH_2CH_2OH$	5.89
32	S	H	O	Me	Bu	5.92
33	S	H	O	Me	$CH_2OCH_2CH_2F$	5.96
34	S	H	O	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	6.07
35	S	H	O	Et	CH_2OCH_2-c-Hex	6.35
36	S	H	S	Et	CH_2OCH_2-c-Hex	6.46
37	S	H	O	Et	CH_2O-*i*-Pr	6.47
38	S	H	O	Me	CH_2OCH_2Me	6.48
39	CH_2	H	O	Et	$CH_2OCH_2CH_2OH$	6.49
40	S	3,5-Me	O	Me	$CH_2OCH_2CH_2OH$	6.59
41	S	3,5-Me	S	Me	$CH_2OCH_2CH_2OH$	6.66
42	S	H	S	Et	CH_2O-*i*-Pr	6.66
43	CH_2	H	O	Et	Bu	6.68
44	S	H	O	Et	$CH_2OCH_2CH_2OH$	6.92
45	S	H	S	Et	$CH_2OCH_2CH_2OH$	6.96
46	S	H	S	c-Pr	CH_2OCH_2Me	7.02
47	S	H	O	Et	$CH_2OCH_2CH_2Ph$	7.02
48	S	H	S	Et	$CH_2OCH_2CH_2Ph$	7.04
49	S	H	O	Me	CH_2OCH_2Ph	7.06
50	S	H	S	Et	$CH_2OCH_2C_6H_4$(4-Me)	7.11
51	S	H	O	*i*-Pr	$CH_2OCH_2CH_2OH$	7.20
52	CH_2	H	O	*i*-Pr	$CH_2OCH_2CH_2OH$	7.20
53	S	H	S	*i*-Pr	$CH_2OCH_2CH_2OH$	7.23
54	S	3,5-Cl	S	Et	$CH_2OCH_2CH_2OH$	7.37
55	CH_2	H	O	*i*-Pr	Bu	7.38
56	CH_2	H	O	Et	CH_2OCH_2Me	7.39
57	S	H	O	Et	CH_2OCH_2Me	7.72
58	S	H	S	*i*-Pr	CH_2OCH_2Me	7.85
59	S	3,5-Me	O	Et	$CH_2OCH_2CH_2OH$	7.89
60	S	3,5-Cl	S	Et	CH_2OCH_2Me	7.89
61	CH_2	3,5-Me	O	Et	$CH_2OCH_2CH_2OH$	7.89
62	S	H	S	Et	$CH_2OCH_2C_6H_4$(4-Cl)	7.92
63	S	H	O	*i*-Pr	CH_2OCH_2Me	7.92
64	S	H	S	Et	CH_2OCH_2Ph	8.11
65	S	3,5-Cl	O	Et	CH_2OCH_2Me	8.13
66	S	3,5-Me	S	Et	CH_2OCH_2Ph	8.16
67	S	H	S	*i*-Pr	CH_2OCH_2Ph	8.17
68	S	H	O	Et	CH_2OCH_2Ph	8.23
69	S	3,5-Me	O	Et	CH_2OCH_2Me	8.27
70	S	3,5-Me	S	*i*-Pr	$CH_2OCH_2CH_2OH$	8.30
71	CH_2	H	O	*i*-Pr	CH_2OCH_2Me	8.38
72	S	3,5-Me	O	Et	CH_2OCH_2Ph	8.49
73	S	3,5-Me	O	*i*-Pr	$CH_2OCH_2CH_2OH$	8.57

Table 20. (Table 4 contd)

no.	R1	R2	X	Y	Z	log(1/EC_{50})
74	S	H	O	*i*-Pr	CH_2OCH_2Ph	8.57
75	CH_2	3,5-Me	O	*i*-Pr	$CH_2OCH_2CH_2OH$	8.57
76	S	4-Me	O	Me	$CH_2OCH_2CH_2OH$	3.66
77	S	3-OH	O	Me	$CH_2OCH_2CH_2OH$	4.09
78	S	3-tBu	O	Me	$CH_2OCH_2CH_2OH$	4.92
79	S	3-I	O	Me	$CH_2OCH_2CH_2OH$	5.00
80	S	H	O	CH═CHPh	$CH_2OCH_2CH_2OH$	5.22
81	S	H	O	Et	CH_2O−c-Hex	5.40
82	S	H	O	CH_2CH═CH_2	$CH_2OCH_2CH_2OH$	5.60
83	S	H	S	Et	CH_2O−c-Hex	5.80
84	S	H	S	Me	$CH_2OCH_2CH_2OH$	6.01
85	CH_2	H	O	Et	CH_2OCH_2OMe	6.60
86	S	H	O	c-Pr	CH_2OCH_2Me	7.00
87	CH_2	H	O	*i*-Pr	CH_2OCH_2OMe	7.28
88	S	H	S	Et	CH_2OCH_2Me	7.59
89	S	3,5-Cl	O	Et	$CH_2OCH_2CH_2OH$	7.85
90	S	3,5-Me	S	Et	$CH_2OCH_2CH_2OH$	8.11
91	S	3,5-Me	S	Et	CH_2OCH_2Me	8.36
92	CH_2	3,5-Me	O	Et	CH_2OCH_2Me	8.80
93	CH_2	3,5-Me	O	*i*-Pr	CH_2OCH_2Me	9.22

[a] 1−75 are in the training set while 76−93 are in the test set.

Table 5. Details of the 18 Descriptors Used To Model HEPT Derivatives

index	descriptor	descriptions
1	$I_{2,4}$	equal to 1 if 2′ and/or 4′ are/is substituted and 0 otherwise
2	$I_{3=5}$	equal to 1 if 3′ and 5′ positions are both substituted by the same substituent and 0 otherwise
3	Y_{iso}	equal to 1 for isopropyl and 0 otherwise
4	Z_{phe}	equal to 1 for phenyl and 0 otherwise
5	R_1	equal to 1 for S atom and 0 otherwise
6	X	equal to 1 for O atom and 0 otherwise
7	R_2_π	π value at the R_2 position
8	R_2_MR	MR value at the R_2 position
9	R_2_F	F value at the R_2 position
10	R_2_R	R value at the R_2 position
11	Y_π	π value at the Y position
12	Y_MR	MR value at the Y position
13	Y_F	F value at the Y position
14	Y_R	R value at the Y position
15	Z_π	π value at the Z position
16	Z_MR	MR value at the Z position
17	Z_F	F value at the Z position
18	Z_R	R value at the Z position

electronic parameters, such as F and R, became available, and absolute MR values were used. To answer the first question, a MLR analysis was first carried out using the truncated descriptor set, {π_3, MR_3', MR_4, MR_5'}, which yielded r^2 of 0.571 and q^2 of 0.539. In addition, similar MLR analyses were performed using the calculated descriptor sets of {π_3, MR_3, MR_4, MR_5}, yielding {r^2, q^2} of {0.240, 0.205}. It is of no surprise that the linear regression results for the nontruncated descriptors are less satisfactory than the truncated ones. Previous QSAR studies indicated that the relationship between MR and activity is nonlinear, and the truncation of this descriptor is an effective way to obtain a better fit when a linear regression method was used.[4] To determine if adding nonlinearity to the correlation would result in a better quality of fit, artificial neural network analyses were carried out using both truncated and calculated descriptor sets yielding {r^2, q^2} as {0.868, 0.652} and {0.845, 0.667}, respectively. Apparently, the results were significantly improved in these ANN runs over the MLR analyses, indicating strong nonlinear dependence of inhibition data on the input descriptors used. To further characterize this nonlinearity, a functional dependence plot[6] of predicted inhibitory activity, pK_i, as a function of MR_3, MR_4, and MR_5 was plotted (Figure 1). All three series of data share the same shape: pK_i values first increase as MR values increase, remain relatively constant in the neighborhood of the maximum, and finally drop in the shape of a parabola as MR values increase further. This suggests that, with larger substituents, the interactions between receptor residues and compounds initially are enhanced, resulting in an increase in activity. However, as substituents become too large, unfavorable steric repulsions between ligand and receptor begin to dominate, resulting in a decrease in activity. As mentioned earlier, Dietrich et al. arbitrarily set the maximum value of MR to be 0.79.[4] Our analysis indicates that this number is a reasonable cutoff value, since this is the vicinity where inhibitory activity begins to drop for MR_3 (see Figure 1). In addition, an approximately linear dependence of pK_i on MR_3 and MR_5 (Figure 1) for descriptor values lower than 0.79 was observed, which explains why the MLR analysis using truncated MR values yielded a good quality of fit. It is interesting to note that the ANN results using calculated

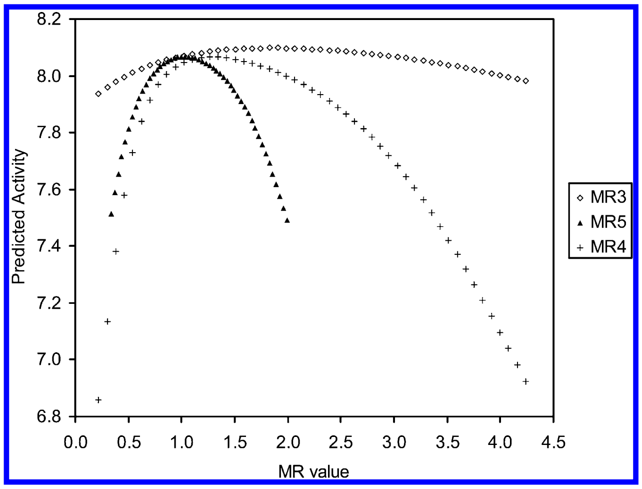


Figure 1. A functional dependence plot of predicted inhibitory activity, log(1/C), as a function of MR3, MR4, and MR5.

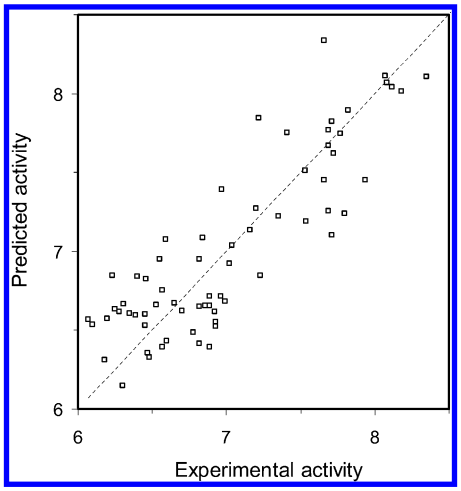


Figure 2. The distribution plot of the cross-validated prediction versus experimental activity data for the best 4-descriptor model for the DHFR inhibitors pyrimidines. The dashed line represents the perfect correlation.

descriptor values are comparable to the ANN results using the truncated descriptor set. This finding is significant since there was no application of expert knowledge involved in the model building process.

The second question is the following: What would be the best model if the absolute MR values were used and additional electronic descriptors {F, R} were available for the building of QSAR models? To address this question, an exhaustive search of all possible four-descriptor models out of a total of twelve descriptors was carried out. The best four-descriptor model {π_3, π_4, F_4, MR_5} from this search did include an electronic parameter. Figure 2 shows a scatter plot of the cross-validated versus the experimental biological activity data for this model. The r^2 and q^2 values for the training set and the leave-one-out cross-validation runs are 0.878 and 0.735, respectively. This result compares favorably with the earlier model based on the original truncated set of descriptors {0.868, 0.652}, suggesting that the electronic attributes may also play an important role in accounting for ligand-protein interactions.

2. QSAR Studies on HEPT Derivatives. *PLS Analysis.* To provide a baseline for comparison with the ANN model, a PLS analysis was carried out on the training set of inhibitory activity and 18 descriptors of 75 derivatives. The best PLS model for calculated descriptors is a two-component model. The r^2 for the training set as well as the leave-one-out cross-validation runs are 0.777 and 0.673, respectively. For the 18 test compounds, the r^2 value is 0.857 for the two-component PLS model.

ANN Analysis. To identify the best QSAR model for HEPT derivatives, an exhaustive search of all possible models of various descriptor numbers was carried out, and the best models were identified for each fixed number of descriptors based on the value of q^2. As can be seen in Figure 3, the internal predictivity first increases as the number of descriptors increases from 2 to 8, remains relatively constant between 8 and 13, and drops slowly afterward. The best 8-descriptor model consists of the following descriptors: $I_{3,5}$, Y_{iso}, Z_{phe}, R_2_R, Y_π, Z_π, Z_MR and Z_R. The relevant statistical numbers for this model are as follows: $r^2 = 0.908$, $q^2 = 0.837$, rms = 0.397. Figure 4 shows a scatter plot of the predicted versus the experimental biological activity data

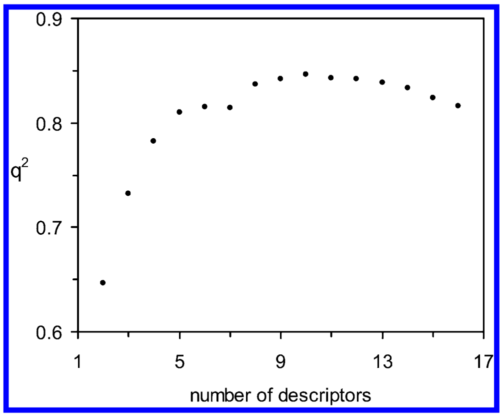


Figure 3. The dependence of q^2 on the number of calculated descriptors.

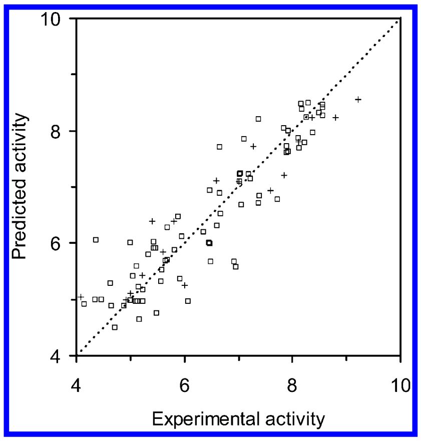


Figure 4. The predicted versus experimental activity data for the cross-validation (shown in squares) and test set runs (shown in crosses) for the best QSAR model for HEPT derivatives. The dashed line represents the perfect correlation.

in the cross-validation as well as the test set runs for this model. The QSAR model also demonstrated good predictivity in the test set: $r^2 = 0.880$, rms = 0.628. Both results are comparable to the best previous QSAR results as summarized in Table 2. These results were achieved by using simple 2D descriptors that have physicochemical meaning. More importantly, the descriptor sets of the best models show consistency with the established SAR. As mentioned in the Introduction, previous QSAR studies have confirmed that I_{iso}, $I_{3,5}$, and Z_{phe} are crucial descriptors to account for inhibitory activity.[16] In addition, it is known that substituents at the Y position should be lipophilic. The fact that Y_π was selected in our QSAR model is consistent with previous findings. It was also known that substituents at the Z position need to be limited in size. In addition, the receptor environment near the Z position is lipophilic.[16] Therefore, it is not surprising that Z_π and Z_MR were also included in our final model.

To validate the statistical significance of the best QSAR models, randomization tests were conducted by arbitrarily shuffling the biological activities of the compounds and then calculating any remaining correlation between the selected descriptors and scrambled activities. This procedure was repeated 50 times with various randomized data sets. The results for calculated descriptor sets are shown in Figure 5. As can be seen, the random test results are well separated from the real data, indicating that the original constructed QSAR models were statistically significant.

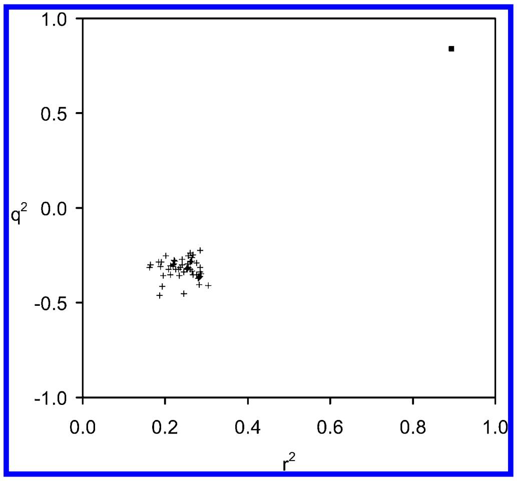


Figure 5. Randomization test result for calculated case. The random test results are shown in crosses, and the real data are shown in squares.

CONCLUSION

The applications of a Hansch substituent constant predictor are demonstrated in QSAR studies of *E. coli* DHFR inhibitors 2,4-diamino-5-(substituted-benzyl)pyrimidines as well as HIV-1 RT inhibitors 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives. When combined with ANN, the predictor allowed us to generate easily interpretable QSAR models of comparable or better predictivity than previous QSAR models.

ACKNOWLEDGMENT

T.-L.C. is grateful to Hoffmann-La Roche for funding her postdoctoral research. We thank Drs. Robert Goodnow and Nora McDonald for helpful discussions and comments on this manuscript.

REFERENCES AND NOTES

(1) Chiu, T. L.; So, S. S. Development of neural network QSPR models for Hansch substituent constants. 1. Method and validations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 147−153.

(2) Ertl, P. World Wide Web-based system for the calculation of substituent parameters and substituent similarity searches. *J. Mol. Graphics Modelling* **1998**, *16*, 11−13.

(3) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.

(4) Dietrich, S. W.; Blaney, J. M.; Reynolds, M. A.; Jow, P. Y.; Hansch, C. Quantitative structure-selectivity relationships. Comparison of the inhibition of *Escherichia coli* and bovine liver dihydrofolate reductase by 5-(substituted-benzyl)-2,4-diaminopyrimidines. *J. Med. Chem.* **1980**, *23*, 1205−1212.

(5) Selassie, C. D.; Li, R. L.; Poe, M.; Hansch, C. On the optimization of hydrophobic and hydrophilic substituent interactions of 2,4-diamino-5-(substituted-benzyl)pyrimidines with dihydrofolate reductase. *J. Med. Chem.* **1991**, *34*, 46−54.

(6) So, S. S.; Richards, W. 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.

(7) Hirst, J. D. Nonlinear quantitative structure−activity relationship for the inhibition of dihydrofolate reductase by pyrimidines. *J. Med. Chem.* **1996**, *39*, 3526−3532.

(8) Li, R. L.; Dietrich, S. W.; Hansch, C. Quantitative structure-selectivity relationships. Comparison of the inhibition of *Escherichia coli* and bovine liver dihydrofolate reductase by 5-(substituted-benzyl)-2,4-diaminopyrimidines. *J. Med. Chem.* **1981**, *24*, 538−544.

(9) Hansch, C.; Li, R.; Blaney, J. M.; Langridge, R. Comparison of the inhibition of *Escherichia coli* and Lactobacillus casei dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl)pyrimidines: quantitative structure−activity relationships, X-ray crystallography, and computer graphics in structure−activity analysis. *J. Med. Chem.* **1982**, *25*, 777−784.

(10) Hirst, J. D.; King, R. D.; Sternberg, M. J. Quantitative structure−activity relationships by neural networks and inductive logic programming. II. The inhibition of dihydrofolate reductase by triazines. *J. Comput. Aided Mol. Des.* **1994**, *8*, 421−432.

(11) King, R. D.; Srinivasan, A. The discovery of indicator variables for QSAR using inductive logic programming. *J. Comput. Aided Mol. Des.* **1997**, *11*, 571−580.

(12) Loukas, Y. L. Adaptive neuro-fuzzy inference system: an instant and architecture-free predictor for improved QSAR studies. *J. Med. Chem.* **2001**, *44*, 2772−2783.

(13) Landavazo, D. G.; Fogel, G. B.; Fogel, D. B. Quantitative structure−activity relationships by evolved neural networks for the inhibition of dihydrofolate reductase by pyrimidines. *Biosystems* **2002**, *65*, 37−47.

(14) Hasegawa, K.; Kimura, T.; Funatsu, K. Nonlinear CoMFA using QPLS as a novel 3D-QSAR approach. *Quant. Struct.-Act. Relat.* **1997**, *16*, 219−223.

(15) Seri-Levy, A.; Salter, R.; West, S.; Richards, W. G. Shape similarity as a single independent variable in QSAR. *Eur. J. Med. Chem.* **1994**, *29*, 687−694.

(16) Gaudio, A. C.; Montanari, C. A. HEPT derivatives as nonnucleoside inhibitors of HIV-1 reverse transcriptase: QSAR studies agree with the crystal structures. *J. Comput. Aided Mol. Des.* **2002**, *16*, 287−295.

(17) Hansch, C.; Zhang, L. QSAR of HIV inhibitors. *Bioorg. Med. Chem. Lett.* **1992**, *2*, 1165−1169.

(18) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I. et al. 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.

(19) Garg, R.; Kurup, A.; Gupta, S. P. Quantitative structure−activity relationship studies on some acyclouridine derivatives acting as anti-HIV-1 drugs. *Quant. Struct.-Act. Relat.* **1997**, *16*, 20−24.

(20) Kireev, D. B.; Chretien, J. R.; Grierson, 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.

(21) Hannongbua, S.; Nivesanond, 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.

(22) 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.

(23) 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.

(24) So, S. S.; Karplus, M. Evolutionary optimization in quantitative structure−activity relationship: an application of genetic neural networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.

(25) So, S. S.; Karplus, M. Genetic neural networks for quantitative structure−activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABAA receptors. *J. Med. Chem.* **1996**, *39*, 5246−5256.

(26) Andrea, T. A.; Kalayeh, H. Applications of neural networks in quantitative structure−activity relation-ships of dihydrofolate reductase inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(27) So, S. S.; van Helden, S. P.; van Geerestein, V. J.; Karplus, M. Quantitative structure−activity relationship studies of progesterone receptor binding steroids. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 762−772.

CI030294I