Nonlinear Quantitative Structure−Activity Relationship for the Inhibition of Dihydrofolate Reductase by Pyrimidines

Jonathan D. Hirst*

Department of Molecular Biology, MB-19, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, California 92037

Received March 11, 1996[⊗]

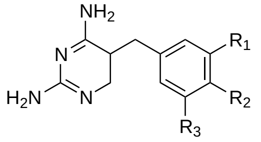
A novel method for quantitative structure−activity relationship (QSAR) analysis is presented. The method, which does not assume any particular functional form for the QSAR, develops nonlinear relationships between parameters describing a set of molecules and the activity of the molecules. For the QSAR of the inhibition of *Escherichia coli* dihydrofolate reductase by 2,4-diamino-5-(substituted benzyl)pyrimidines, the method compares favorably to other nonlinear methods. Cross-validation trials demonstrate that the predictive ability is as accurate as other methods, and the method is simpler and faster than neural network and machine-learning methods. Consequently, its implementation is much easier, and interpretation of the generated QSAR is more straightforward.

Introduction

When the absence of detailed structural data about a drug−substrate complex precludes the application of structure-based drug design, insights about binding often come from the development of a quantitative structure−activity relationship (QSAR). Recent years have seen a flurry of novel QSAR algorithms that seek to go beyond the linear and quadratic models introduced and popularized by Hansch.[1,2] In particular, there have been many applications of neural networks to QSAR analysis.[3−13] Machine-learning techniques have also been developed.[14−17] One of the primary goals of these studies (which have been reviewed recently[18]) is to capture subtle relationships that elude more traditional techniques. Thus, particular dependencies, such as linear or quadratic, are usually not assumed, and, instead, more complex nonlinear relationships are developed. The development of these nonlinear relationships involves, in the case of neural networks, searching a weight-space that is of much higher dimensionality than the space of the QSAR problem itself. For example, for a typical QSAR there might be five parameters considered to be of interest, and the corresponding neural network developed to model this QSAR would typically have 30 or more weights. Similarly, some machine-learning techniques also require the search of high-dimensional spaces, which are too large to search exhaustively.

In this work, we present an algorithm which develops nonlinear relationships in a well-defined, simple fashion, working in the parameter space of interest. If a molecule is characterized by its activity and p properties, and there are N molecules in the data set, the algorithm generates a surface of activity as a function of p in a $(p + 1)$-dimensional space from the N points. In this particular application, we have been able to reduce the number of properties, so that $p = 2$, and the generated surface may be readily visualized.

We have applied the method to the QSAR of the inhibition of dihydrofolate reductase (DHFR) by pyrimidines. DHFR plays a key role in the synthesis of DNA. DHFR from different organisms may be inhibited by the same molecule to differing degrees, and this differential inhibition makes DHFR inhibitors candidates as antibacterial agents. The inhibition of *Escherichia coli* DHFR by trimethoprim and its analogues, benzylpyrimidines with substituents R_1, R_2, and R_3 at the 3-, 4-, and 5-positions (**I**), has been studied extensively by X-ray crystallography,[19,20] and activity data have been measured for 74 related molecules.[21−24] These data have been used as a test case for a variety of QSAR algorithms, including neural networks and machine-learning applications.[4,8,15,21,22,25] These previous studies are used as a benchmark for the work presented here.



I

In this work, we describe a method for generating nonlinear QSARs. The predictive accuracy of the method is assessed using a cross-validation trial on a comparatively large and well-studied data set. This is compared to the reported accuracies for several other methods. The simplicity of the method is self-evident and allows facile interpretation of the generated QSAR. An obvious, but previously unremarked upon, feature of the data set allows a reduction in the dimensionality of the problem. This kind of feature may be present in other QSAR data. The focus of this work, however, is not on variable selection or reduction, both of which are key problems, but on the efficient generation of nonlinear QSARs given the variables.

The advent of combinatorial chemistry technologies[26,27] provides an additional impetus for the development of fast and accurate QSARs, with the goal of directing the robotic synthesis of compounds in real time. The ability to synthesize and assay large numbers of compounds will be best exploited if one can generate and utilize meaningful QSARs with sufficient speed and accuracy. The necessity of speed and the large numbers

* Tel: (619)-554-3736. Fax: (619)-554-6688. E-mail: jhirst@scripps.edu.

[⊗] Abstract published in *Advance ACS Abstracts,* August 1, 1996.

S0022-2623(96)00197-5 CCC: $12.00

Table 1. Properties and Activities of the 74 Pyrimidines Used in This Study[a]

		activity					
no.	substituents	measd	pred	MR$_5'$	MR$_3'$	MR$_4$	π_3
			Set 1				
11	4-F	6.35	6.23	0.10	0.10	0.09	0.00
31	4-$NHCOCH_3$	6.89	6.37	0.10	0.10	1.49	0.00
34	3-Br	6.96	6.92	0.10	0.79	0.10	0.86
42	3,5-$(OCH_3)_2$, 4-$(CH_2)_2OCH_3$	8.35	8.26	0.79	0.79	1.93	0.00
20	3-OCH_2CONH_2	6.57	6.80	0.10	0.79	0.10	−1.37
24	3-CH_3	6.70	6.98	0.10	0.57	0.10	0.52
30	4-$O(CH_2)_3CH_3$	6.89	6.40	0.10	0.10	2.17	0.00
23	3-Cl	6.65	6.99	0.10	0.60	0.10	0.67
37	3-CF_3	7.02	6.99	0.10	0.79	0.10	0.50
08	3-CH_2OH	6.28	6.80	0.10	0.72	0.10	−1.03
39	3-I	7.23	6.81	0.10	0.79	0.10	1.12
			Set 2				
16	3-OH	6.47	6.42	0.10	0.28	0.10	0.28
54	3,5-Cl_2, 4-NH_2	8.87	8.07	0.60	0.60	0.54	0.71
43	3,5-$(OCH_3)_2$	8.38	8.35	0.79	0.79	0.10	−0.02
19	3-$CH_2O(CH_2)_3CH_3$	6.55	6.91	0.10	0.79	0.10	1.30
22	3-CH_2OCH_3	6.59	6.65	0.10	0.79	0.10	−0.78
10	3,5-$(CH_2OH)_2$	6.31	8.35	0.72	0.72	0.10	−1.03
41	3,4-$(OCH_3)_2$	7.72	6.86	0.10	0.79	0.79	−0.02
33	3-OCH_3	6.93	6.71	0.10	0.79	0.10	−0.02
35	3-NO_2, 4-$NHCOCH_3$	6.97	7.03	0.10	0.74	1.49	−0.28
01	3,5-$(OH)_2$	3.04	6.39	0.28	0.28	0.10	−0.67
09	4-NH_2	6.30	6.35	0.10	0.10	0.54	0.00
			Set 3				
14	4-Cl	6.45	6.24	0.10	0.10	0.60	0.10
03	4-$O(CH_2)_5CH_3$	6.07	6.50	0.10	0.10	3.07	0.00
55	3-Cl, 4-NH_2, 5-CH_3	8.87	7.98	0.57	0.60	0.54	0.71
06	3-F	6.23	6.17	0.10	0.09	0.10	0.23
04	H	6.18	6.18	0.10	0.10	0.10	0.00
47	3,5-$(OCH_3)_2$, 4-$O(CH_2)_5CH_3$	7.87	8.21	0.79	0.79	0.79	−0.02
28	3-$O(CH_2)_3CH_3$	6.82	7.04	0.10	0.79	0.10	1.55
25	4-$N(CH_3)_2$	6.78	6.40	0.10	0.10	1.56	0.00
27	4-OCH_3	6.82	6.20	0.10	0.10	0.79	0.00
50	3,5-$(OCH_3)_2$, 4-CH_3	8.57	8.34	0.79	0.79	0.57	−0.02
07	3-$O(CH_2)_7CH_3$	6.25	6.84	0.10	0.79	0.10	3.71
			Set 4				
51	3,5-I_2, 4-OCH_3	8.82	8.61	0.79	0.79	0.79	1.12
49	3,5-$(OCH_3)_2$, 4-$OCH_2C_6H_5$	8.42	8.20	0.79	0.79	0.79	−0.02
53	3,5-Br_2, 4-NH_2	8.85	7.82	0.79	0.79	0.54	0.86
46	3,5-$(CH_3)_2$, 4-OCH_3	7.74	8.50	0.57	0.57	0.79	0.56
12	3-$O(CH_2)_6CH_3$	6.39	6.74	0.10	0.79	0.10	3.17
21	4-OCF_3	6.57	6.25	0.10	0.10	0.79	0.00
44	3,4,5-$(OCH_3)_3$	8.87	8.20	0.79	0.79	0.79	−0.02
52	3,5-I_2, 4-OH	8.82	8.68	0.79	0.79	0.28	1.12
38	3,4-$(COH_2CH_2OCH_3)_2$	7.22	7.05	0.10	0.79	1.93	−0.40
15	3,4-$(OH)_2$	6.46	6.21	0.10	0.28	0.28	−0.67
18	3-$OCH_2CH_2OCH_3$	6.53	6.81	0.10	0.79	0.10	−0.40
			Set 5				
36	3-$OCH_2C_6H_5$	6.99	6.78	0.10	0.79	0.10	1.66
45	3,5-$(CH_3)_2$, 4-OH	7.56	8.62	0.57	0.57	0.28	0.56
26	4-Br	6.82	6.36	0.10	0.10	0.89	0.00
32	3-OSO_2CH_3	6.92	6.69	0.10	0.79	0.10	−0.88
02	4-$O(CH_2)_6CH_3$	5.60	6.27	0.10	0.10	0.79	0.00
17	4-CH_3	6.48	6.48	0.10	0.10	0.57	0.00
40	3-CF_3, 4-OCH_3	7.69	6.91	0.10	0.50	0.79	0.88
13	4-$OCH_2CH_2OCH_3$	6.40	6.39	0.10	0.10	0.93	0.00
48	3,5-$(OCH_3)_2$, 4-$O(CH_2)_7CH_3$	7.87	8.36	0.79	0.79	0.79	−0.02
29	3-$O(CH_2)_5CH_3$	6.86	6.60	0.10	0.79	0.10	2.63
05	4-NO_2	6.20	6.23	0.10	0.10	0.74	0.00
			Set 6				
56	4-OH	6.45	6.30	0.10	0.10	0.29	0.00
57	4-OSO_2CH_3	6.60	6.48	0.10	0.10	1.70	0.00
58	3-OH, 4-OCH_3	6.84	6.17	0.10	0.28	0.79	−0.67
59	4-$OCH_2C_6H_5$	6.89	6.46	0.10	0.10	3.17	0.00
60	4-C_6H_5	6.93	6.52	0.10	0.10	2.54	0.00
61	3,5-$(CH_3)_2$	7.04	8.52	0.57	0.57	0.10	0.56
62	3,4-$(OCH_2O)_2$	7.13	6.28	0.10	0.45	0.45	0.00
63	3-$O(CH_3)_7OCH_3$, 4-OCH_3	7.16	6.60	0.10	0.79	0.88	3.71
64	3,5-$(OCH_3)_2$, 4-$O(CH_3)_7OCH_3$	7.20	8.29	0.79	0.79	3.97	−0.02
65	3,5-OC_3H_7	7.41	8.60	0.79	0.79	0.10	1.05
66	3-OCH_3, 4-$CH_2C_6H_5$	7.53	7.12	0.10	0.79	3.17	−0.02
67	3-OCH_3, 4-OH	7.54	6.85	0.10	0.79	0.28	−0.02
68	3-$OCH_2C_6H_5$, 4-OCH_3	7.66	7.21	0.10	0.79	0.79	1.27
69	3,5-$(OCH_3)_2$, 4-$N(CH_3)_2$	7.71	8.30	0.79	0.79	1.56	−0.02
70	3-OCH_3, 4-$O(CH_2)_2OCH_3$	7.77	7.12	0.10	0.79	1.93	−0.02
71	3-OSO_2CH_3, 4-OCH_3	7.80	6.81	0.10	0.79	0.79	−0.88
72	3,4,5-$(CH_2CH_3)_3$	7.82	8.48	0.79	0.79	1.03	0.86
73	3-OCH_3, 4-OSO_2CH_3	7.94	7.09	0.10	0.79	1.70	−0.02
74	3,5-$(OCH_3)_2$, 4-Br	8.18	8.35	0.79	0.79	0.89	−0.02

[a] The index numbers are those given in previous work.[8]

of compounds may preclude the time-consuming computation of 3-dimensional properties of the molecules, and so initially we focus on more traditional QSAR parameters.

Method

A QSAR relates p properties to activity. This defines a (p + 1)-dimensional space. A QSAR is a surface in this space; linear QSARs are planes in this space, and quadratic QSARs are parabolic surfaces. We generate a nonlinear surface through the following prescription. A grid is defined in the p-dimensional property space. The fineness of the grid is arbitrary; in this study 32 grid points in each dimension were used. At each grid point, g, the activity, $A_{[g]}$, is a distance-weighted sum of the activities of the molecules in the data set:

$$A_{[g]} = W\sum_{i}^{N}\frac{A_i}{d_i} \qquad (1)$$

where

$$d_i = \sum_{j}^{p}(p_{ij} - p_{gi})^2 \qquad (2)$$

and

$$W = \sum_{i}^{N} d_i \qquad (3)$$

W is a normalization constant, N is the number of molecules, d is the square of the Euclidean distance, p is the number of properties, p_{ij} is the value of the jth property of molecule i, and p_{gj} is the value of the jth property at grid point g. In the case when the grid point is at the identical location of a molecule, that molecule is excluded from the sum, because otherwise d, on the right-hand side of eq 1, is zero and $A_{[g]}$ would not be defined. In fact, any molecule very close (defined as $d < 0.05$) to a grid point is excluded from the sum to calculate the activity at that grid point. While this is computationally convenient, it also provides a mechanism for generalization, in that the computed surface is not overly sensitive to the idiosyncrasies of individual data points.

Thus, the activity at any point in property space is assumed to be similar to the activity of nearby molecules. The distance-weighting means that only molecules near to a particular point influence the computed activity at that point. Functional forms other than the one in eq 1 could also be used, thus altering the form of the distance-weighting. We show later that the predictive accuracy is relatively insensitive to the functional form, as long as the underlying mathematical model that molecules with similar properties have similar activities is retained. We find that the details of similarity, in terms of the precise values of the weights in the distance-weighted sum, do not overly influence the performance of the algorithm. Clearly, the concept of the distance-weighted sum is not novel; however, to the best of our knowledge, its application to QSAR analysis is. Combined with a reduction in the dimensionality of the problem, the method provides a powerful tool for QSAR analysis.

The data used in this study are 74 2,4-diamino-5-(substituted benzyl)pyrimdines. The properties of the substituents of these molecules and their activities were taken from the literature.[21-24] The activity data have been reported as $\log(1/K_i)$, where K_i is the experimentally measured inhibition constant. These data come from two different laboratories where different assay methods were employed. Ideally, the activity data would all come from the same assay performed in the same laboratory. However, as the focus of this work is to assess the performance of a novel method, it is essential to use data which have been studied by several nonlinear methods. Given that these data have been used in several previous studies,[8,15,25] the identical data were used in this study to provide meaningful comparisons. Furthermore, the two different assays are in acceptable agreement on a number of overlapping data points. The chemical properties of the substituents at the 3-, 4-, and 5- positions are the hydrophobic parameter, π, which is derived from the ratio of the partition coefficients in 1-octanol and water, and the molar refractivity, MR, which is related to the size and the polarizability of the substituent. These MR values were rescaled in the earlier work of Hansch and co-workers,[21,28] so no further rescaling has been introduced here. In a distance-weighted approach, such as ours, clearly the scaling of different dimensions is important. We do not need to address this issue for the data in this study, and a more general discussion of rescaling is not within the scope of the work presented here.

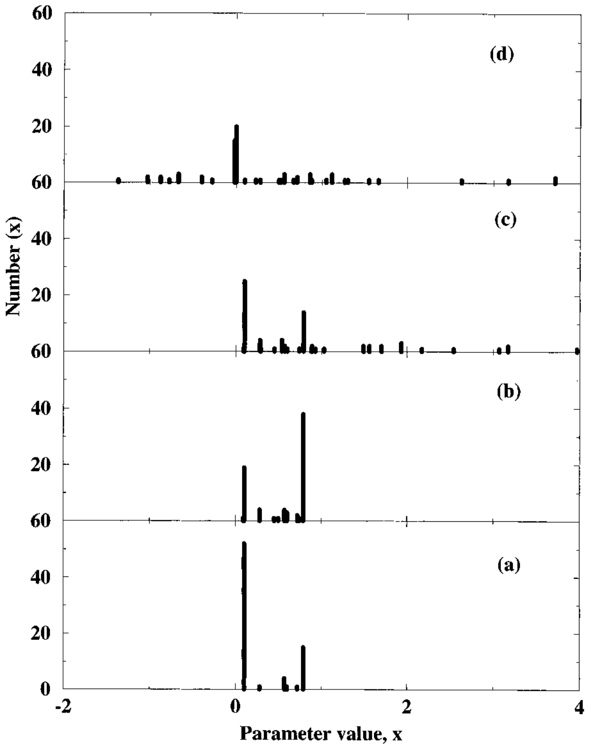


Figure 1. Distribution of the parameters used to describe 74 pyrimidines: (a) MR_5'—the truncated molar refractivity of the substituent at the 5-position, (b) MR_3'—the truncated molar refractivity of the substituent at the 3-position, (c) MR_4—the molar refractivity of the substituent at the 4-position, and (d) π_3—the hydrophobicity of the substituent at the 3-position.

These data were divided into five sets of 11 for a cross-validation trial and a separate set of 19, as shown in Table 1. Fivefold cross-validation was performed to assess predictive accuracy. Set 6 was only used as an additional test set. In the first split of the data, set 1 was the test set and sets 2−5 comprised the training set; in the second split, set 2 was the test set and sets 1 and 3−5 comprised the training set, and similarly for the other splits in the cross-validation trial. Each molecule is described by four parameters: MR_5', MR_3', MR_4, and π_3. The prime indicates a truncation of the range of the parameter as introduced by Hansch[21,28] (and discussed later in this section). For consistency, molecules that are singly substituted at the meta-position are treated as 3-substituted molecules, with MR_3' and π_3 assigned according to the substituent and MR_5' assigned for hydrogen.

As presented in Table 1, the QSAR is a 4-dimensional problem. While a 5-dimensional surface can be generated with ease using the method described above, a 3-dimensional surface is clearly more easily visualized and potentially more insightful. Furthermore, working in a lower dimensional space has computational and statistical advantages. These benefits and the identification of collinear parameters are motivations for techniques such as principal components analysis,[29] in which a smaller number of new parameters are developed from linear combinations of the original parameters. However, in the data under study here, a simpler means of reducing the dimensionality was suggested by inspection of the data.

Table 2. Distribution of Pairs of Parameters Used To Describe the 74 Pyrimidines

parameters	no. diff pairs	most common pair (no. occurrences)	2nd most common pair (no. occurrences)	3rd most common pair (no. occurrences)
$\{MR_5', MR_3'\}$	16	{0.10, 0.79} (23)	{0.10, 0.10} (19)	{0.79, 0.79} (15)
$\{MR_5', MR_4\}$	38	{0.10, 0.10} (20)	{0.10, 0.79} (8)	{0.79, 0.79} (5)
$\{MR_5', \pi_3\}$	35	{0.10, 0.00} (19)	{0.79, −0.02} (9)	{0.10, −0.02} (6)
$\{MR_3', MR_4\}$	43	{0.79, 0.10} (16)	{0.79, 0.79} (8)	{0.79, 1.93} (3)
$\{MR_3', \pi_3\}$	30	{0.10, 0.00} (18)	{0.79, −0.02} (15)	{0.79, 0.12} (3)
$\{MR_4, \pi_3\}$	65	{0.79, −0.02} (5)	{0.79, 0.00} (3)	{0.54, 0.71} (2)

In Figure 1, the distributions of the properties of the substituents are shown. It can be seen that molecules in the data set predominantly have only one of two values for the parameters MR_3' and MR_5', either 0.10 or 0.79. This arises for two reasons. Firstly, only a small number of different substituents appear in the data: most common are hydrogen and the methoxy group. Secondly, these parameters are truncated MR values, which means that any value above 0.79 has been reported as 0.79. This arose from an earlier QSAR analysis, in which it was concluded that larger substituents with greater values did not appear to be more effective inhibitors. As is evident from Table 2, in which all the pairwise distributions of parameters are summarized, the data fall into three classes: molecules with small 3- and 5-substituents ($MR_3' = MR_5' = 0.10$), molecules with a large 3-substituted and a small 5-substituent ($MR_3' = 0.79$, $MR_5' = 0.10$), and molecules with bulky 3- and 5-substituents ($MR_3' = MR_5' = 0.79$). To accommodate the small number of molecules that do not exactly adhere to the above classification, we have defined the three classes as $MR_3' < 0.5$, $MR_5' < 0.5$; $MR_3' \geq 0.5$, $MR_5' < 0.5$; and $MR_3' \geq 0.5$, $MR_5' \geq 0.5$.

Thus, we have reduced the 4-dimensional problem to three 2-dimensional problems. While this is indeed convenient, it is not merely a matter of convenience. There are very little data in the region between the extremes of MR_3' and MR_5', and interpolation in this region by any QSAR algorithm would have to be viewed with skepticism. While this observation may be obvious, it does not appear to have been considered in previous analyses of these data.

In the cross-validation trials, the training data were used to calculate three surfaces of activity as a function of MR_4 and π_3, one surface for each of the three combinations of MR_3' and MR_5' described above. The activity of the test data was predicted as the activity at the nearest grid point to the test data point, for the surface corresponding to the MR_3' and MR_5' values of the test molecule.

Results

Table 3 shows that the predictive accuracy of the method, as assessed by cross-validation and on an independent test set of 19 molecules, compares favorably to other methods. The cross-validated Spearman rank correlation coefficient between the predicted and measured activities is 0.68. The method performs as well as any in the cross-validation trial, which gives a more reliable estimate of predictive accuracy than the independent trial, because the number of data points is greater (55 compared to 19). On the independent test set, two of the other methods perform better, but the difference is not statistically significant. For the training data, the mean cross-validated Spearman rank correlation coefficient between the predicted and measured activities is 0.80. This is lower than for the other methods, but it is performance on unseen test data not on training data that is of primary interest.

Table 3. Comparison of the Predictive Accuracy of the Nonlinear QSAR with the Accuracies of Other Methods, as Reported in the Literature[25]

method	mean cross-validation performance[a] (σ)		mean independent test set performance (σ)
	training data	test data	
multiple linear regression[b]	0.89 (0.05)	0.65 (0.10)	0.51 (0.15)
nearest neighbor[c]	1.00 (0.00)	0.55 (0.17)	0.50 (0.07)
neural network[d]	0.89 (0.04)	0.68 (0.12)	0.63 (0.18)
M5[e]	0.82 (0.09)	0.62 (0.26)	0.59 (0.07)
CART[f]	0.98 (0.00)	0.50 (0.29)	0.54 (0.16)
GOLEM[g]	0.95 (0.01)	0.68 (0.11)	0.74 (0.10)
this work: nonlinear QSAR	0.80 (0.02)	0.68 (0.16)	0.59 (0.01)

[a] Accuracy is measured by the Spearman rank correlation coefficient between the actual and predicted activities of the test data. All accuracies were calculated by 5-fold cross-validation on the same data. The means and standard deviations were computed from the five cross-validation trials. [b] The multiple linear regression was performed using the variables and their squares. [c] The nearest neighbor algorithm assigns to the test molecule the activity of the nearest training set molecule. [d] The neural network used back-propagation learning with a speed-up algorithm.[8] [e] M5 is a machine-learning method based on a combination of regression trees and instance-based learning.[30] [f] Classification and regression trees, CART, is a collection of binary decision tree-growing algorithms.[31] [g] GOLEM is a machine-learning method based on inductive logic programming.[8,15,32]

In the cross-validation trial, two molecules are predicted to have much higher activities than their measured activities. These are the 3,5-$(OH)_2$- and 3,5-$(CH_2OH)_2$-substituted molecules, whose predicted activities are 6.39 and 8.35, respectively, compared to their measured activities of 3.04 and 6.31. Both these molecules have been identified as anomalies in previous studies.[21,28] It appears that the presence of two hydroxyl groups in the active site is unfavorable, much more so than one or two methoxy groups, and that the description of substituents by MR and π variables does not fully capture the pertinent interactions. Excluding these two outliers, the mean absolute error in the predicted activity of the other molecules in the cross-validation trial is 0.34.

In Figure 2 we show surfaces generated from the entire data set. A number of features are apparent. Activity tends to increase with the molar refractivity of the 3- and 5-substituents. If both substituents have low molar refractivities (Figure 2a), the mean activity is 6.16 (standard deviation, $\sigma = 1.36$; size of sample, $n = 25$); if MR_3' is large and MR_5' is small (Figure 2b), the mean activity is 7.04 ($\sigma = 0.49$; $n = 28$); if both substituents have large molar refractivities (Figure 2c), the mean activity increases to 8.07 ($\sigma = 0.70$; $n = 21$). The dependence of the activity on π_3 and MR_4 varies

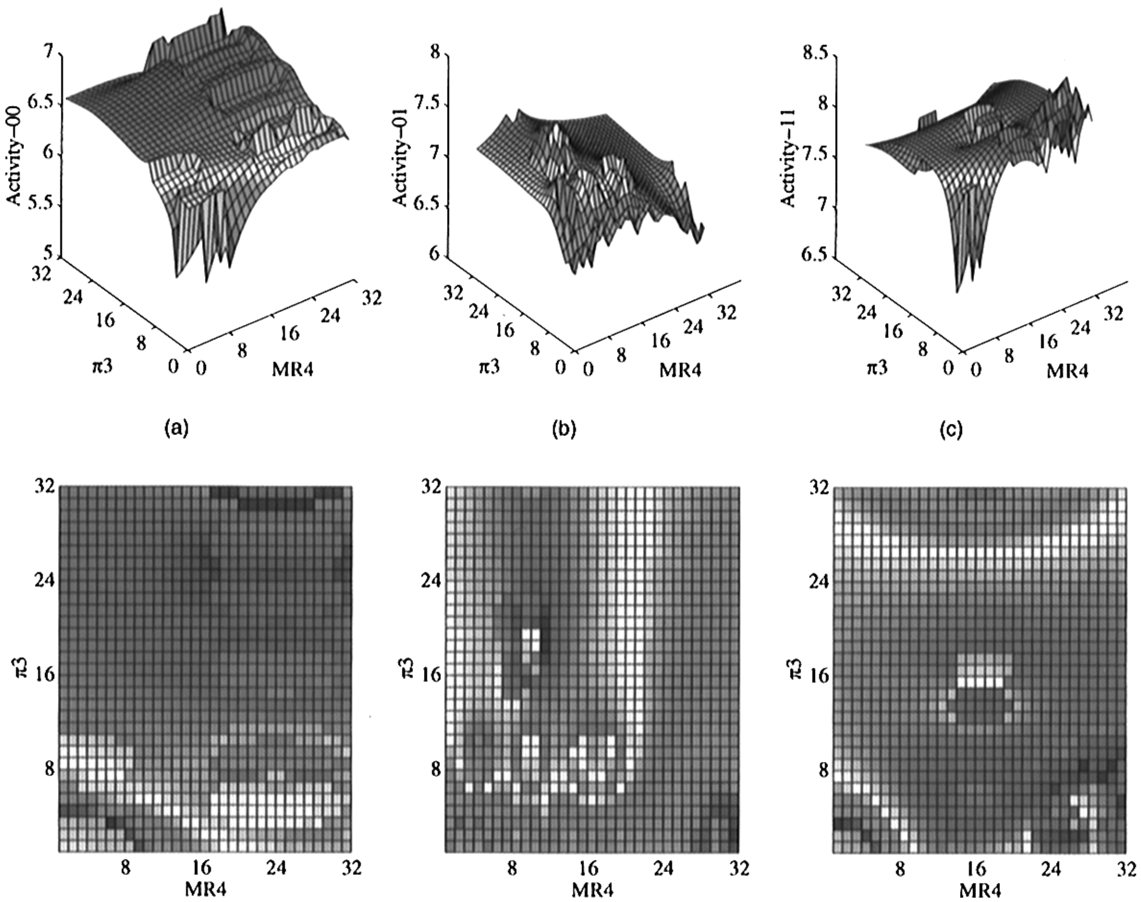


Figure 2. Nonlinear QSARs for (a) molecules with low MR_3' and MR_5', (b) molecules with high MR_3' and low MR_5', and (c) molecule with high MR_3' and MR_5'. The respective activities are denoted as activity-00, activity-01, and activity-11. In each plot, the activities are color-coded, with red indicating high activity (within the range of the particular plot) and blue indicating low activity. Projections of the 3-dimensional plots are shown in the lower portion of the figure. The plots are shown as 32 × 32 grids, as generated by the method. In each plot, the ranges of π_3 and MR_4 have been divided into 32 equal divisions. The actual range of π_3 is (a) −0.67 to 0.28, (b) −1.37 to 3.71, and (c) −1.03 to 1.12. The actual range of MR_4 is (a) 0.09 to 3.17, (b) 0.10 to 3.17, and (c) 0.10 to 3.97.

significantly for these three different subsets of molecules. This is an indication of the importance of cross-terms in the QSAR. For a particular substituent, the properties required to give high activity depend on the properties of the other substituents. This is in contrast to a typical linear QSAR, where each property can usually be independently optimized.

For molecules with 3- and 5-substituents with small molar refractivities, activity increases fairly uniformly with π_3 and MR_4, although a high MR_4 value on its own is not sufficient for high activity. Low π_3 gives low activity. The highest activity for these molecules is $\log(1/K_i) = 7.13$. Molecules with high MR_3' but small MR_5' have a different QSAR. Low π_3 or high MR_4 leads to low activity. The most active molecules have low to moderate MR_4 and moderate to high π_3. The most active molecules are those in which both the 3- and 5-substituents have high molar refractivities. Of these, the most active molecules have low π_3 values and high MR_4 values.

As alluded to earlier, many functional forms may be used in the distance-weighted sum in eq 1. We have explored a number of these through a generalization of eq 1:

$$A_{[g]} = W_{ck}\sum_{i}^{N} A_i\left(\frac{1}{cd^k} + c - \frac{1}{c}\right) \qquad (4)$$

where W_{ck} is the modified normalization constant, k changes the exponent of the distance, so $k = 1$ gives the squared distance, $k = 0.5$ gives the distance itself, and c defines the slope of the weighting curve, i.e., how rapidly molecules away from the grid point are essentially ignored. Equation 1 is recovered if $c = k = 1$. Table 4 shows that the predictive ability of the method is relatively insensitive to these parameters. As might be expected, the predictive ability decreases as more extreme functional forms are adopted for the distance-weighting, where either too many or too few near neighbors are allowed to contribute with significant weights.

Table 4. Dependence of the Predictive Accuracy on Different Functional Forms[a]

	c				
k	0.1	0.5	1	2	5
0.25	0.49	0.55	0.68	0.70	0.70
0.5	0.47	0.56	0.69	0.70	0.70
1	0.59	0.46	0.68	0.66	0.65
2	0.67	0.58	0.66	0.63	0.61
4	0.57	0.61	0.63	0.61	0.62

[a] The 5-fold cross-validation test set performance is given as a function of the parameters c and k in eq 4.

Conclusions

The method we have presented for nonlinear QSAR analysis has the following advantages: (i) generality—no particular dependence of activity on properties is assumed; (ii) simplicity—especially compared to neural networks or machine-learning algorithms; (iii) accuracy—it is as accurate as other methods, as assessed by cross-validation; (iv) speed—the 5-fold cross-validation trial took 3 cpu s on a SGI Indy workstation (with a 134 MHz MIPS R4600 chip); (v) low dimensionality—it works in the space of the QSAR parameters; and (vi) ease of interpretation—the 3-dimensional color-coded surfaces are readily visualized.

Furthermore, the results are insensitive to small changes in the definition of the method. There are a couple of mild qualifications to these conclusions. Firstly, while the nonlinear QSAR is accurate, it is not more accurate than the other methods it was compared to, and the QSAR of the inhibition of DHFR by pyrimidines is still not fully understood. Nevertheless, our study has provided further insight into the QSAR. Studies of other QSARs are currently underway to provide further assessment of the method. Secondly, the ease of interpretation provided by the 3-dimensional surfaces is substantially reduced in higher dimensions, although one could look at projections on a 3-dimensional space.

The primary aim of this work was to present a method for developing nonlinear QSARs. The data in this study were chosen as a well-studied test case. However, as well as illustrating the performance of the nonlinear QSAR, our study has detected important features of the data. In particular, we note that the data fall into three distinct classes of molecules with different mean activities. This allowed the QSAR to be reduced to three 2-dimensional problems, which may be readily visualized.

For very large dimensional problems, the algorithm will require some modification. The computational effort for the generation of large-dimensional surfaces grows exponentially, although activities can be computed at points corresponding to test data, without computing the entire surface. However, careful selection of parameters and the judicious reduction of the parameter space by, for example, principal component analysis should in most QSARs obviate the need to work in a very high dimensional space. This is, in general, desirable for computational efficiency, statistical significance, and ease of interpretation.

Acknowledgment. The Human Frontiers Science Program is gratefully acknowledged for financial support in the form of a long-term fellowship. Professor Charles L. Brooks III is thanked for critical reading of the manuscript and for many stimulation discussions.

References

(1) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with Hammett substitution constants and partition coefficients. *Nature* **1962**, *194*, 178−180.

(2) Hansch, C. A quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

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

(4) 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.

(5) Villemin, D.; Cherqaoui, D.; Cense, J. M. Neural network studies: quantitative structure-activity relationship of mutagenic aromatic nitro compounds. *J. Chim. Phys.* **1993**, *90*, 1505−1519.

(6) Ajay. A unified framework for using neural networks to build QSARs. *J. Med. Chem.* **1993**, *36*, 3565−3571.

(7) King, R. D.; Hirst, J. D.; Sternberg, M. J. E. New approaches to QSAR: Neural networks and machine learning. *Perspect. Drug Discovery Des.* **1993**, *1*, 279−290.

(8) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure-activity relationships by neural networks and inductive logic programming I. The inhibition of dihydrofolate reductase by pyrimidines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 405−420.

(9) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. 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.

(10) Tetko, I. G.; 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.

(11) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758−3767.

(12) Ajay. On better generalization by combining two or more models: A quantitative structure-activity relationship example using neural networks. *Chem. Intell. Lab. Systems* **1994**, *24*, 19−30.

(13) Maddalena, D. J.; Johnston, G. A. R. Prediction of receptor properties an binding affinity of ligands to benzodiazepine/GABA$_A$ receptors using artificial neural networks. *J. Med. Chem.* **1995**, *38*, 715−724.

(14) Bolis, G.; Pace, L. D.; Fabrocini, F. A machine learning approach to computer-aided molecular design. *J. Comput.-Aided. Mol. Des.* **1991**, *5*, 617−628.

(15) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. E. Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 11322−11326.

(16) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M. J. E. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 438−442.

(17) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: A shape-based machine learning tool for drug design. *J. Comput.-Aided Drug Des.* **1994**, *8*, 635−652.

(18) Jackson, R. C. Update on computer-aided drug design. *Curr. Opin. Biotech.* **1995**, *6*, 646−651.

(19) Champness, J. N.; Stammers, D. K.; Beddell, C. R. Crystallographic investigation of the cooperative interaction between trimethoprim, reduced cofactor and dihydrofolate reductase. *FEBS Letts.* **1986**, *199*, 61−67.

(20) Matthews, D. A.; Bolin, J. T.; Burridge, J. M.; Filman, D. J.; Volz, K. W.; Kaufman, B. T.; Beddell, C. R.; Champness, J. N.; Stammers, D. K.; Kraut, J. Refined crystal structures of Escherichia coli and chicken liver dihydrofolate reductase containing bound trimethoprim. *J. Biol. Chem.* **1985**, *260*, 381−391.

(21) 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.

(22) 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.

(23) Roth, B.; Aig, E.; Rauckman, B. S.; Srelitz, J. Z.; Phillips, A. P.; Ferone, R.; Bushby, S. R. M.; Siegel, C. W. 2,4-Diamino-5-benzylpyrimidines and analogues as antibacterial agents. 5. 3′,5′-Dimethoxy-4′-substituted-benzyl analogues of trimethoprim. *J. Med. Chem.* **1981**, *24*, 933−941.

(24) Roth, B.; Rauckman, B. S.; Ferone, R.; Baccanari, D. P.; Champness, J. N.; Hyde, R. M. 2,4-Diamino-5-benzylpyrimidines as antibacterial agents. 7. Analysis of the effect of 3,5-dialkyl substituent size and shape on binding to four different dihydrofolate reductase enzymes. *J. Med. Chem.* **1987**, *30*, 348−356.

(25) King, R. D.; Hirst, J. D.; Sternberg, M. J. E. Comparison of artificial intelligence methods for modeling pharmaceutical QSARs. *Appl. Artif. Intell.* **1995**, *9*, 213−234.

(26) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(27) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(28) Hansch, C.; Li, R.-I.; 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.

(29) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.

(30) Quinlan, J. Induction of decision trees. *Machine Learning* **1986**, *1*, 81−106.

(31) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, 1984.

(32) Muggleton, S.; Feng, C. Efficient induction of logic programs. In *Proceedings of the First Conference on Algorithmic Learning Theory*; Arikawa, S., Goto, S., Ohsuga, S., Yokomori, T., Eds.; Jpn. Soc. Artificial Intelligence: Tokyo, 1990; pp 368−381.

JM960197Z