0.5 H, H-4''), 3.70 (m, 0.5 H, H-1''), 3.00-2.50 (complex m, 1 H, H-5'' of cis isomer cis to S and H-5'' of trans isomer cis to S), 2.29 (m, 0.5 H, H-5'' of trans isomer trans to S), 1.93 (m, 0.5 H, H-5'' of cis isomer trans to S), 1.61 (acetonide CH_3), 1.44 (s, 9 H, Boc-CH_3), 1.41 (s, 3 H, acetonide CH_3); IR (neat) 3329 (N—H), 2888 (aliphatic C—H), 1679 (C=O), 1463, 1385 (C=C) cm^{-1}. Anal. ($C_{23}H_{32}N_6O_5S$) C, H, N.

***S*-(5'-Deoxy-5'-adenosyl)-1-amino-4-mercapto-2-cyclopentene (14).** A 0.150-g portion of 13 was dissolved in 3 mL of 88% formic acid and allowed to stir at room temperature for 2 days. The reaction mixture was then diluted to 25 mL with water, and the aqueous layer was extracted with three 25-mL portions of ether and lyophilized to yield a pale yellow solid. The solid was then purified by chromatography on silica gel ($CHCl_3$/MeOH/NH_4OH 9:2:1); fractions containing the product (R_f = 0.52) were pooled, 50 mL of water was added, and the mixture was concentrated on a rotary evaporator until the pH of the aqueous layer was neutral. The aqueous solution was lyophilized to afford pure 14 (0.093 g, 86.2% yield) as a white solid: ^1H NMR (CD_3OD) δ 8.31 (s, 1 H, H-2), 8.21 (s, 1 H, H-8), 6.17 (d, 1 H, H-1'), 5.97 (m, 1 H, H-2''), 5.80 (m, 1 H, H-3''), 5.02 (m, 1 H, H-2'), 4.79 (m, 1 H, H-3'), 4.32 (m, 1 H, H-4'), 4.18 (m, 1 H, H-1''), 4.06 (m, 1 H, H-4''), 3.30 (m, 2 H, H-5'), 2.97 (complex m, 2 H, H-5''); IR (KBr) 3336 (N—H), 3192 (O—H), 2917 (aliphatic C—H) cm^{-1}. Anal. ($C_{15}H_{20}N_6O_3S$) C, H, N.

***S*-(5'-Deoxy-5'-adenosyl)-1-ammonio-4-(methylsulfonio)-2-cyclopentene Disulfate (3).** Compound 14 (0.093 g, 0.00026 mol) was dissolved in 2 mL of a 50:50 mixture of formic acid and acetic acid along with 0.148 g (0.065 mL, 0.001 mol) of iodomethane. A 0.108-g (0.00052 mol) portion of silver perchlorate dissolved in 1.08 mL of 50:50 formic acid/acetic acid (10% w/v) was then added. The reaction mixture was allowed to stir at room temperature overnight, after which the yellow precipitate was removed by centrifugation. The clear solution was diluted to 50 mL with water and extracted with three 25-mL portions of ether, and the aqueous layer was lyophilized to afford a light yellow solid. The solid residue was chromatographed on silica gel (butanol/acetic acid/water 1:1:1), and the product-containing fractions were pooled and diluted to 50 mL with water. The aqueous layer was washed with 25 mL of ether and lyophilized to give a white solid. The solid was dissolved in 0.1 N H_2SO_4, and ethanol was added to precipitate the product as a white solid (0.107 g, 71.6%): ^1H NMR (CD_3OD) δ 8.49 (s, 1 H, H-2), 8.47 (s, 1 H, H-8), 6.46 (m, 1 H, H-2''), 6.40 (m, 1 H, H-3''), 6.12 (d, 1 H, H-1'), 5.08 (m, 1 H, H-2'), 4.75 (m, 1 H, H-3'), 4.58 (m, 1 H, H-4'), 3.91 (complex m, 2 H, H-1'' and H-4''), 2.90 (complex m, 2 H, H-5''), 2.82 (s, 3 H, CH_3); IR (KBr) 3536, 3403 (N—H, O—H), 2935 (aliphatic C—H) cm^{-1}. Anal. ($C_{16}H_{24}N_6O_3S \cdot 2HSO_4 \cdot 0.5EtOH$) C, H, N.

Enzyme Purification. AdoMet-DC is isolated from *Escherichia coli* using a modification of the methylglyoxal bis(guanylhydrazone) (MGBG) Sepharose affinity column procedure of Anton and Kutny.[21] The column is prepared by incubating MGBG with epoxy-activated Sepharose 4B at pH 11 as described. *E. coli* (3/4 log phase, Grain Processing Co., Ames, IA) are lysed in 5 volumes of 10 mM Tris-HCl, 0.5 mM EDTA, and 0.5 mM dithiothreitol, pH 8.0, by a single pass through a French press. A 5% solution of streptomycin sulfate is then added to give a final concentration of 1%, and the cell debris is removed by centrifugation at 20000*g* for 2 h. AdoMet-DC is allowed to adsorb to the gel by stirring the gel and the lysate supernatant together for 1 h after bringing the $MgCl_2$ concentration to 10 mM. Binding is considered complete when residual AdoMet-DC activity in the supernatant is determined to be 1-3% of the original value. The gel is then packed into a column and washed (20 mM Tris-HCl, 10 mM $MgCl_2$, 0.6 M KCl, 0.5 mM EDTA, and 0.5 mM dithiothreitol, pH 8.0) until UV absorption reaches baseline. AdoMet-DC is then eluted using 20 mM potassium phosphate, 0.6 M KCl, 0.5 mM EDTA, and 0.5 mM dithiothreitol, pH 7.0, and the fractions of highest activity are pooled and concentrated (Amicon ultrafiltration cell, PM-30 membrane). Protein is measured by the method of Bradford.[22] With this method enzyme purity is greater than 90%, and the specific activity is determined to be 0.80 μmol/min per mg of protein at 37 °C.

Enzyme Assay. AdoMet-DC activity is monitored by following the evolution of $[^{14}C]CO_2$ from *S*-adenosyl-L-[carboxy-^{14}C]-methionine using a modification of the method of Markham.[23] Each reaction mixture contains 50 μg of AdoMet-DC, 40 μL of *S*-adenosyl-L-[carboxy-^{14}C]methionine (0.9 mCi/mmol, 20 μM final concentration) in 62.5 mM Tris-HCl/100 mM $MgSO_4$, pH 7.4, with a final volume of 2 mL. Radiolabeled CO_2 is trapped on a filter disk in a vial cap soaked with 40 μL of hyamine. After 15 min the reaction is quenched, and the disk is placed in a scintillation vial with 10 mL of scintillation cocktail and counted (counting efficiency 95% or greater). Each data point represents the average of two determinations, which in each case differ by less than 5%.

Acknowledgment. Financial support through a research grant from the Elsa U. Pardee Foundation is gratefully acknowledged. We are also indebted to Mr. Kirk A. Douglas and Mr. Kelvin Grant for excellent technical assistance.

(21) Anton, D. L.; Kutny, R. Escherichia coli S-Adenosylmethionine Decarboxylase: Subunit Structure, Reductive Amination and NH_2-Terminal Sequences. *J. Biol. Chem.* **1987**, *262*, 2817-2822.

(22) Bradford, M. M. A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding. *Anal. Biochem.* **1976**, *72*, 248-254.

(23) Markham, G. D.; Tabor, C. W.; Tabor, H. S-Adenosylmethionine Decarboxylase of *Escherichia coli*. *J. Biol. Chem.* **1982**, *257*, 12063-12068.

Application of Neural Networks: Quantitative Structure-Activity Relationships of the Derivatives of 2,4-Diamino-5-(substituted-benzyl)pyrimidines as DHFR Inhibitors

Sung-Sau So and W. Graham Richards*

Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, United Kingdom. *Received March 5, 1992*

A comparative study of quantitative structure-activity relationships involving diaminopyrimidines as DHFR inhibitors using regression analysis and the neural-network approach suggests that the neural network can outperform traditional methods. The technique permits the highlighting the functional form of those parameters which have an influence on the biological activity.

Introduction

The formulation of quantitative structure-activity relationships (QSAR) has had a momentous impact upon medicinal chemistry for the past 30 years. Hansch demonstrated that the biological activities of drug molecules can be correlated by a linear combination of the physicochemical parameters of the corresponding drug. Since then there have been many attempts to include cross-product terms in the regression analysis, but this only added complexity to the study and resulted in no significant im-

0022-2623/92/1835-3201$03.00/0

provement.

Recently there has been growing interest in the application of neural networks in the field of QSAR. It has been demonstrated that this new technique is often superior to the traditional Hansch approach. The key strength of the neural networks is that with the presence of hidden layers, neural networks are able to perform nonlinear mapping of the physicochemical parameters to the corresponding biological activity implicity.[1] This is especially true for the networks with a large number of nodes in the hidden layer, and some impressive results have been obtained.[2,3] However, there is a danger of "overfitting", that is to say the number of variables under the control of the neural networks may exceed the number of data points that are needed to describe the hypersurface. In such cases the neural network simply memorizes the entire data set and it is effectively a look-up table. It is doubtful that the network would be able to extract relevant correlation of the input patterns and give meaningful interpretation of other unknown examples. We need to stress that the purpose of QSAR is to understand the forces governing the activity of a particular class of compound, and to assist drug design. A look-up table will not aid medicinal chemists in the design of new drugs. What is needed is a system that is able to provide reasonable predictions for the compounds which are previously unknown.

There are two advantages of adopting networks with a small number of hidden units. Firstly, the efficiency of each node increases and consequently the time of the computer simulation is significantly reduced. Secondly, and more importantly, the network can generalize the input patterns better, and this results in superior predictive power. However, caution is again needed. Just as a two-layer perceptron cannot solve the XOR problem (a perceptron may be regarded as a neural network with no hidden units),[4] a network with insufficient hidden units will not be able to extract all the relevant correlation between physicochemical parameters and biological activity. The analysis will collapse at the point of training and again no reliable predictions may be obtained.

Thus the neural networks programmer must be cautious about the architecture of the network being constructed. While the numbers of nodes in the input and output layers are likely to be predetermined by the nature of experimental data, the freedom is really the number of hidden units. It has been suggested that a ratio, ρ, plays a crucial role in determining the number of hidden units being employed.[1,5,6] The definition of ρ is

$$\rho = \frac{\text{number of data point in the training set}}{\text{number of variables controlled by the network}}$$

The number of variables is simply the sum of the number of connections in the network and the number of biases. A three-layer back-propagation network with I inputs units, H hidden units, and O output units will have $H(I + O)$ connections and $H + O$ biases. The total number of adjustable parameters is therefore $H(I + O) + H + O$. The range $1.8 < \rho < 2.2$ has been suggested as a guideline of acceptable ρ values.[1] 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.

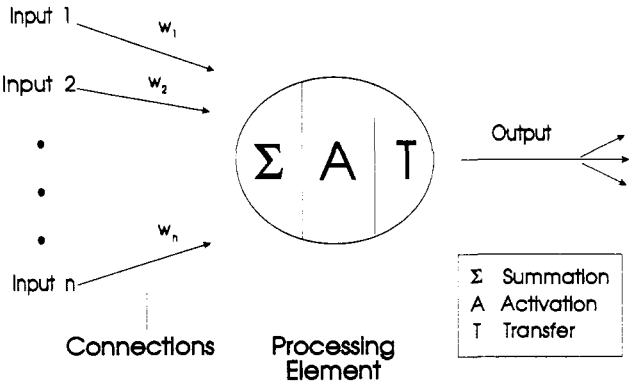

Figure 1. Schematic representation of a computer neuron.

The ρ ratio has a significant impact upon the design of neural-network architecture. It is now possible to make a sensible choice of the number of hidden nodes without too much worry about either the overfitting effect or the generalization problem. Nevertheless, the suggested range of $1.8 < \rho < 2.2$ is perhaps empirical, and is also likely to be implementation-dependent. Some correlation may already exist in the training patterns so that the "effective" number of data points may in fact be smaller than anticipated. In this paper we use this range as a rough guideline.

A Monte Carlo algorithm is also implemented in the neural-network simulator.[7] This is used to search for a suitable set of initial starting weights. The purpose of this is to allow searches to be made in a larger weight-space and consequently the results obtained should be better statistically.

Theory

Neural Networks. Neural networks, also known as parallel distributed processing models, neurocomputers, or connectionist models, are computer-based simulations of living nervous systems. They consists of a large number of simple processing elements, analogous to biological neurons, which are extensively interconnected to form a highly parallel computer.[8,9] Neural networks are only first-order approximations of the brain. The essential components of neural networks are the processing elements, the connections, and the topology (Figure 1).

Processing Elements. The processing element is a simplified model of a neuron. Physiologists know that there are at least 150 processes being performed in biological neurons, yet only four of these functions are being emulated in the computer neurons. They are (a) input and output (I/O) function, which evaluates input signals from the neurons of the previous layer, determining the strength

(1) 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.
(2) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Structure–Activity Relationships. *J. Med. Chem.* 1990, *33*, 905–908.
(3) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure–Activity Relationship Analysis. *J. Med. Chem.* 1990, *33*, 2583–2590.
(4) Minsky, M.; Papert, S. *Perceptrons: An Introduction to Computational Geometry*; MIT Press: Cambridge, MA, 1969.
(5) Livingstone, D. J.; Salt, D. W. Regression Analysis for QSAR Using Neural Networks. Submitted.
(6) Manallack, D. T.; Livingstone, D. J. Chance Effects Using (Artificial) Neural Networks for Data Analysis. Submitted.
(7) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vettering, W. T. *Numerical Recipes in C*; Cambridge University Press: Cambridge, 1988.
(8) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley Publishing Co.: Reading, MA, 1991.
(9) McCord Nelson, M.; Illingworth, W. T. *A Practical Guide to Neural Nets*, Addison-Wesley Publishing Co.: Reading, MA, 1991.

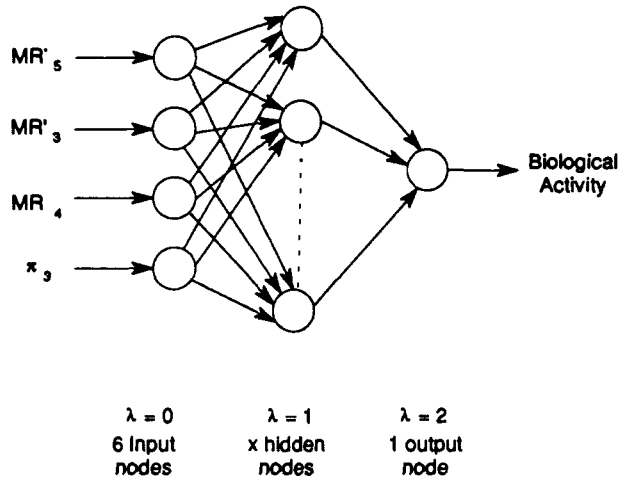

Figure 2. A multiple-layer network with a 4-x-1 configuration as used in our study.

of each input, and passes the output signal to the neurons of the next layer; (b) summation function, which calculates a total for the combined input signals according to the equation

$$\text{net input}_{\lambda,i} = \sum_j \text{weight}_{ij}\text{output}_{\lambda-1,j}$$

where i and j are units in layers λ and $\lambda - 1$, respectively; (c) activation function, which allows the outputs to vary with respect to time. The result of summation is passed to this function before it is input to the transfer function, and typically it is an identity function for most current implementations; and (d) transfer function, which maps the summed input to an output value. The most commonly used transfer function is the logistic function

$$\text{output}_{\lambda,j} = f(\text{net input}_{\lambda j}) = \frac{1}{1 + \exp(-\text{net input}_{\lambda j})}$$

It is a continuous and differentiable approximation of a threshold function, and has values close to zero or unity over most of its domain.

Connections. Analogous to the synaptic strengths of biological neurons, each input of the processing element is associated with a relative weight and it affects the impact of that input. This makes some input more important than others in the way they combine to produce an impulse. The physical rationale is that positive weights correspond to excitatory connections and negative weights inhibitory connections. In a simple network, the weights of the connections are the only parameters which are adjusted during the training process. Thus the knowledge contained in the system is stored in the strengths of these connections rather than in the processing elements themselves.

Topology. Many different network topologies are possible: they can be feed-forward nets or feedback nets, single layer or multiple layers, fully connected or partially connected. For the bulk of practical applications, it is found that the multiple-layer feed-forward nets are the most useful, because they are relatively simple to assemble and present fewer problems in training. An example of this type of network is shown in Figure 2.

Training. Training is the way a neural network learns. The methods can be divided into two types. In the unsupervised method the input patterns are presented once into the network and the network settles and converges to a final state. In the supervised method, the error between the actual output of the net and the target output are computed, and on this basis the strengths of the connections are modified. This process is iterated until the weights of the connections are optimized such that the overall error is minimized. At present the unsupervised learning is not well-understood and is still subject of much research interest. On the other hand, supervised learning has been successfully implemented in practical applications.

Back-propagation is a widely used supervised learning method for multiple-layer nets, which seems to be best adapted for solving pattern recognition problems. This technique is the most widely used generalization of the δ rule, and the procedure involves two phases.

The forward phase occurs when the input is presented and propagated forward through the network to compute an output value for each processing element based on its current set of weights.

$$\text{net input}_{\lambda,i} = \sum_j \text{weight}_{ij}\text{output}_{\lambda-1,j}$$

The backward phase is a recurring difference computation performed in a backward direction. The error, δ, for neurons in the output layer is given by

$$\delta_{\lambda,i} = (\text{target output}_{\lambda,i} - \text{actual output}_{\lambda,i})f'(\text{net input}_{\lambda,i})$$

where $f'(x)$ is the first derivative of the transfer function (and hence the requirement of it being continuous and differentiable). For the units in the hidden layer, a specific target value is unknown and it is computed in terms of the errors in the units in the next layer forward

$$\delta_{\lambda,i} = \sum_j (\delta_{\lambda+1,j}\text{weight}_{ji})f'(\text{net input}_{\lambda,i})$$

The weight between the input unit i and the output unit j is then modified according to the equation

$$\Delta\text{weight}_{ji} = \eta\delta_{\lambda,j}\text{output}_{\lambda-1,i}$$

where η is an empirical parameter known as the learning rate. Theoretically, η needs to be infinitesimally small for true gradient descent of error, but in practice, it typically takes values from 0.1 to 1.0, and is being gradually reduced during the training process.[10] A problem often occurs in the training process. Commonly the system gets stuck in a local minimum and fails to reach the global minimum state. To rectify this, researchers often add a momentum term, which takes into considerations of past weight changes, in their calculation of weight adjustments

$$\Delta_{n+1}\text{weight}_{ji} = \eta\delta_{\lambda,j}\text{output}_{\lambda-1,i} + \alpha\Delta_n\text{weight}_{ji}$$

where α is again an empirical parameter similar to η. This is an attempt to bump the system past the barriers in the temporary pockets. Adding random noise (which ultimately decays to zero) to the input patterns helps to avoid local minima, it also makes the network more robust to noisy data. Additionally, if the training examples are chosen in random order, it also makes the path through weight-space stochastic, allowing wider exploration of the hypersurface.

To summarize the learning process, an input is fed into the network in order to calculate the error with respect to the desired target. This error is then used to compute the weight corrections layer by layer, backward through the net. The supervised training process is repeated until the error for all of the training set is minimized. Typically it involves thousands of iterations. After training the network is fully operational.

(10) Elrod, D. W.; Maggiora, G. M. Applications of Neural Networks in Chemistry. 1. Prediction of Electrophilic Aromatic Substitution Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*:4, 477–484.

Table I.[a] Structures, Physicochemical Parameters, and Observed DHFR Inhibitory Activities of the Pyrimidines Congeners

no.	X	MR'_5	MR'_3	MR_4	π_3	activity	no.	X	MR'_5	MR'_3	MR_4	π_3	activity
1	4-$O(CH_2)_5CH_3$	0.10	0.10	3.07	0.00	6.07	35	3-OSO_2CH_3	0.79	0.10	0.10	0.00	6.92
2	4-$O(CH_2)_6CH_3$	0.10	0.10	3.52	0.00	6.10	36	3-OCH_3	0.10	0.79	0.10	0.04	6.93
3	H	0.10	0.10	0.10	0.00	6.18	37	4-C_6H_5	0.10	0.10	2.54	0.00	6.93
4	4-NO_2	0.10	0.10	0.74	0.00	6.20	38	3-Br	0.10	0.79	0.10	0.86	6.96
5	3-F	0.10	0.09	0.10	0.23	6.23	39	3-NO_2, 4-$NHCOCH_3$	0.74	0.10	1.49	0.00	6.97
6	3-$O(CH_2)_7CH_3$	0.10	0.79	0.10	3.79	6.25	40	3-$OCH_2C_6H_5$	0.10	0.79	0.10	1.56	6.99
7	3-CH_2OH	0.72	0.10	0.10	0.00	6.28	41	3-CF_3	0.10	0.50	0.10	0.88	7.02
8	4-NH_2	0.10	0.10	0.54	0.00	6.30	42	3,5-$(CH_3)_2$	0.57	0.57	0.10	0.56	7.04
9	3,5-$(CH_2OH)_2$	0.72	0.72	0.10	-1.03	6.31	43	3,4-OCH_2O	0.45	0.10	0.45	0.00	7.13
10	4-F	0.10	0.10	0.09	0.00	6.35	44	3-$O(CH_2)_7CH_3$, 4-OCH_3	0.10	0.79	0.88	3.69	7.16
11	3-$O(CH_2)_6CH_3$	0.10	0.79	0.10	3.23	6.39	45	3,5-$(OCH_3)_2$, 4-$O(CH_2)_7CH_3$	0.79	0.79	3.97	0.00	7.20
12	4-$OCH_2CH_2OCH_3$	0.10	0.10	0.93	0.00	6.40	46	3,4-$(OCH_2CH_2OCH_3)_2$	0.79	0.10	1.93	0.00	7.22
13	4-OH	0.10	0.10	0.29	0.00	6.45	47	3-I	0.10	0.79	0.10	1.12	7.23
14	4-Cl	0.10	0.10	0.60	0.00	6.45	48	3-OCH_2CH_3, 4-$OCH_2C_6H_5$	0.10	0.79	3.17	0.38	7.35
15	3,4-$(OH)_2$	0.29	0.10	0.29	0.00	6.46	49	3,5-$(OC_3H_7)_2$	0.79	0.79	0.10	1.05	7.41
16	3-OH	0.29	0.10	0.10	0.00	6.47	50	3-OCH_3, 4-$OCH_2C_6H_5$	0.79	0.10	3.17	0.00	7.53
17	4-CH_3	0.10	0.10	0.57	0.00	6.48	51	3-OCH_3, 4-OH	0.79	0.10	0.29	0.00	7.54
18	3-$OCH_2CH_2OCH_3$	0.79	0.10	0.10	0.00	6.53	52	3,5-$(OCH_2CH_3)_2$, 4-pyrryl	0.79	0.79	1.95	0.38	7.66
19	3-$CH_2O(CH_2)_3CH_3$	0.10	0.79	0.10	1.30	6.55	53	3-$OCH_2C_6H_5$, 4-OCH_3	0.10	0.79	0.79	1.27	7.66
20	3-OCH_2CONH_2	0.79	0.10	0.10	0.00	6.57	54	3,5-$(OCH_2CH_3)_2$	0.79	0.79	0.10	0.47	7.69
21	4-OCF_3	0.10	0.10	0.79	0.00	6.57	55	3-OC_2H_5, 5-OC_3H_7	0.79	0.79	0.10	1.05	7.69
22	3-CH_2OCH_3	0.79	0.10	0.10	0.00	6.59	56	3-CF_3, 4-OCH_3	0.10	0.50	0.79	0.87	7.69
23	4-OSO_2CH_3	0.10	0.10	1.70	0.00	6.60	57	3,5-$(OCH_3)_2$, 4-$N(CH_3)_2$	0.79	0.79	1.56	0.00	7.71
24	3-Cl	0.10	0.60	0.10	0.67	6.65	58	3,5-$(OCH_3)_2$	0.79	0.79	0.10	0.00	7.71
25	3-CH_3	0.10	0.57	0.10	0.52	6.70	59	3,4-$(OCH_3)_2$	0.79	0.10	0.79	0.00	7.72
26	4-$N(CH_3)_2$	0.10	0.10	1.56	0.00	6.78	60	3-OCH_3, 4-$OCH_2CH_2OCH_3$	0.79	0.10	1.93	0.00	7.77
27	3-$O(CH_2)_3CH_3$	0.10	0.79	0.10	1.55	6.82	61	3-OSO_2CH_3, 4-OCH_3	0.79	0.10	0.79	0.00	7.80
28	4-OCH_3	0.10	0.10	0.79	0.00	6.82	62	3,4,5-$(CH_2CH_3)_3$	0.79	0.79	1.03	0.86	7.82
29	4-Br	0.10	0.10	0.89	0.00	6.82	63	3-OCH_3, 4-OSO_2CH_3	0.79	0.10	1.70	0.00	7.94
30	3-OH, 4-OCH_3	0.29	0.10	0.79	0.00	6.84	64	3,5-$(OCH_3)_2$, 4-SCH_3	0.79	0.79	1.38	0.00	8.07
31	3-$O(CH_2)_5CH_3$	0.10	0.79	0.10	2.63	6.86	65	3,4,5-$(OCH_3)_3$	0.79	0.79	0.79	0.00	8.08
32	4-$NHCOCH_3$	0.10	0.10	1.49	0.00	6.89	66	3,5-$(OCH_3)_2$, 4-$C(CH_3)=CH_2$	0.79	0.79	1.56	0.00	8.12
33	4-$O(CH_2)_3CH_3$	0.10	0.10	2.17	0.00	6.89	67	3,5-$(OCH_3)_2$, 4-Br	0.79	0.79	0.89	0.00	8.18
34	4-$OCH_2C_6H_5$	0.10	0.10	3.17	0.00	6.89	68	3,5-$(OCH_3)_2$, 4-$O(CH_2)_2OCH_3$	0.79	0.79	1.93	0.00	8.35

[a] The data were taken from ref 12.

Monte Carlo Algorithm. This is a minimization algorithm inspired by the Boltzmann probability distribution, and it is sometimes known as the Metropolis algorithm. The function

$$\text{probability}(E) = \exp\left(\frac{-E}{kT}\right)$$

is the probability of finding a system with energy E. It is important to realize that even at low temperature there is still a finite chance of finding a system at high energy. An overview of this algorithm is as follows.

In a simulation, a configuration of the particles of the system is generated randomly and its energy is calculated. Another random configuration is generated and again the new energy is evaluated. If $E_{new} < E_{old}$, the new configuration is accepted as the starting point immediately. If $E_{new} > E_{old}$, there is a $\exp -(E_{new} - E_{old}/kT)$ probability of acceptance. Under this general scheme, a system may go either uphill or downhill in energy, although the latter process is more likely. The temperature of the system is gradually reduced in the process and the downhill step becomes progressively dominant. It is hoped that this will eventually lead to a global minimum energy state.

In an analogous way, a set of weights was randomly generated and the corresponding error of the input patterns evaluated. The probability of accepting the new random set of weights is again dependent upon a similar exponential factor. The system is believed to be stuck at a "reasonable" error minimum when the number of rejected steps exceeds an user-defined threshold. Neural-network techniques then take over and continue the minimization in the error.

The merit of the Monte Carlo algorithm is its speed, especially when the working system is very large. This algorithm does not work on a definite error surface. It means the problem of a local minimum is largely bypassed.

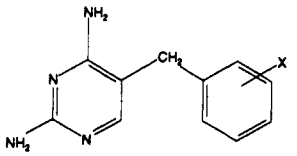

Figure 3. Structure of 2,4-diamino-5-(substituted-benzyl)pyrimidines.

Results

We have performed a QSAR study for the inhibition of DHFR by 2,4-diamino-5-(substituted-benzyl)pyrimidines (Figure 3) using the data in Table I. This particular set of compounds has been extensively studied by Hansch et al. and is ideal for the purpose of comparison.[11,12] The neural network was simulated using a computer program written in the C programming language and run on a variety of platforms, including 386/387 and 486 personal computers and Silicon Graphics and SUN workstations. The input data were normalized to give values between 0.0 and 1.0. Training continued until there was no further decrease in overall error after a period of 50000 cycles. The average training time for each run was about 2 h. Three-layer neural networks, with four input units and one

(11) Hansch, C.; Li, R.-L.; 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.

(12) 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.

Table II. Identities of Compounds in the Training Set and the Test Set

Training Set (49 examples)										
1	2	3	4	5	7	8	10	11	12	13
15	16	17	19	20	23	24	25	26	29	30
31	32	33	35	36	37	39	40	42	43	45
46	48	50	51	53	54	55	56	58	59	61
64	65	66	67	68						
Test Set (19 examples)										
6	9	14	18	21	22	27	28	34	38	41
44	47	49	52	57	60	62	63			

Table III. Comparison of Residual Variance and Rank Correlation Coefficient with Different ρ Values

		training set		testing set	
configuration	ρ	RV	SRCC	RV	SRCC
4-3-1	2.58	0.0364	0.92	0.208	0.59
4-4-1	1.96	0.0239	0.96	0.187	0.74
4-5-1	1.58	0.0183	0.96	0.230	0.67
4-6-1	1.32	0.0150	0.96	0.290	0.61
4-7-1	1.14	0.0126	0.97	0.323	0.40

output unit, were simulated in all cases.

In this paper the quality of QSAR was assessed by two statistical variables: the residual variance (RV)[5] and the Spearman rank correlation coefficient (SRCC).[13] They were defined by the following expressions:

$$RV = \frac{\sum(\text{activity}_{\text{observed}} - \text{activity}_{\text{predicted}})^2}{\text{number of compounds} - 1}$$

$$SRCC = 1 - \frac{6\sum(\text{rank}_{\text{observed}} - \text{rank}_{\text{predicted}})^2}{\text{number of compounds}[(\text{number of compounds})^2 - 1]}$$

High-quality QSAR work should have RV close to zero and SRCC close to unity.

Comparison of Neural-Network- and Regression-Generated Biological Activity Surface. As mentioned the number of hidden nodes is an important factor determining the network's performance. It was desirable to establish a network which generalized the input patterns rather than merely memorizing them. A preliminary study was conducted to determine an appropriate number of hidden units. The 68 compounds were divided into two data sets as indicated by Table II. The first set, which comprised 49 compounds, served as the training set and the remaining 19 were used to give guidance to the accuracy of the trained networks. Five networks were constructed with configurations of 4-x-1, where x = 3–7. Each was simulated at least five times. The results of the simulations which gave the smallest RVs for the training set in each configuration are shown in Table III.

The values of RVs of the training set and the test set were plotted against ρ, and are illustrated in Figure 4a,b. The plots are in accordance with expectation. The RVs of the training set decrease fairly linearly with ρ, and on extrapolation this line seems to pass through the origin. This is consistent with the earlier hypothesis that, given enough adjustable parameters ($\rho \rightarrow 0$), the network would be able to map the complete data set (RV $\rightarrow$ 0). The RVs of the test set, on the other hand, show a different trend. In accordance to the findings of a earlier paper by Andrea et al.,[1] it is a nonlinear function of ρ and has an upward concave shape. It seems that the minimum was within the range of $1.8 < \rho < 2.3$, but the differences in the test set

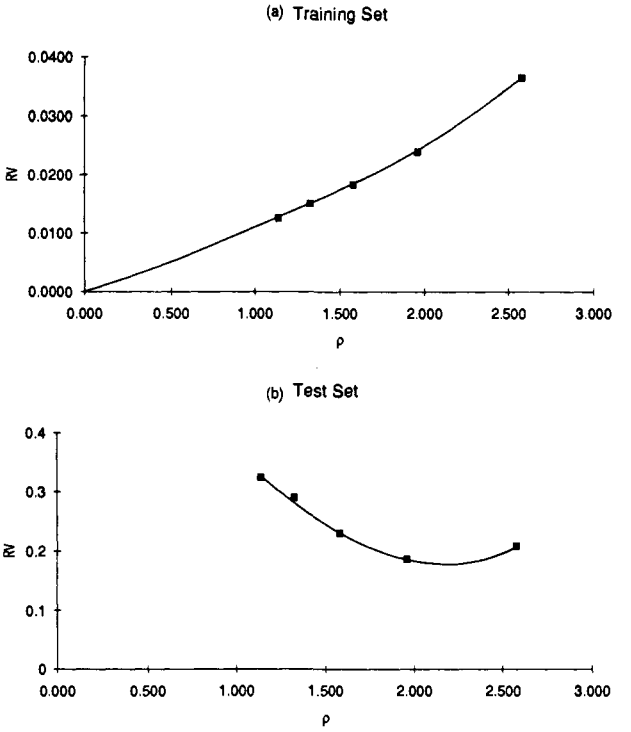

Figure 4. Residual variance of the training set and the test set.

RVs in this range are so small that it is insensitive to ρ within this range. In the light of designing a network which would give reliable predictions, a neural network with configuration giving similar ρ values to the ones corresponding to smallest test set variances was constructed to train all the 68 compounds.

A neural network with the characteristics listed below was constructed for the task:

configuration	number of examples	ρ	η	α	initial range of weights and biases
4-6-1	68	1.83	0.25	0.90	−1.00 to +1.00

The results of QSAR done by this neural network and by regression analysis were listed in Table IV, and comparisons of residual variances, rank correlation, and the number of outliers are shown in Table V. It is clear that neural network outperforms regression analysis and provides superior mapping of physicochemical parameters to biological activities. The ρ value of 1.83 suggests that the network would not be able to memorize the data set and thus the mapping was believed to be an impressive result of generalization.

Prediction of Biological Activities. Having established a configuration of neural network for predictive purpose, a cross-validation procedure was carried out.[14] In this process one compound was removed from the data set, and the remaining 67 compounds served as the training set to a network with a 4-6-1 configuration. After training, the parameters of the compound unknown to the network were put into the network and the predicted biological activity of this compound was evaluated. This procedure was repeated 68 times and the predicted activities of the entire data set were obtained. The results are shown in last column of Table IV. The cross-validated r^2,[14] RV, and SRCC of the predicted set are 0.724, 0.102, and 0.84, respectively. It is estimated that 95% of the observed ac-

(13) Freund, J. E. *Modern Elementary Statistics*; Prentice-Hall: Englewood Cliffs, NJ, 1988.

(14) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* 1988, *110*, 5959.

Table IV. Calculated Activities of the Pyrimidines by Neural Network and Regression Analysis

no.	observed activity	activity calculated: neural network	activity calculated: regression analysis[a]	activity predicted, neural network[b]	no.	observed activity	activity calculated: neural network	activity calculated: regression analysis[a]	activity predicted, neural network[b]
1	6.07	6.52	6.55	6.69	35	6.92	6.66	6.86	6.61
2	6.10	6.20	6.27	6.34	36	6.93	6.86	6.77	6.91
3	6.18	6.27	6.21	6.18	37	6.93	6.83	6.84	6.69
4	6.20	6.52	6.60	6.61	38	6.96	7.00	6.98	6.97
5	6.23	6.28	6.29	6.30	39	6.97	7.05	7.44	7.63
6	6.25	6.39	6.47	6.42	40	6.99	6.87	6.90	6.85
7	6.28	6.26	6.79	6.58	41	7.02	6.84	6.72	6.75
8	6.30	6.44	6.50	6.42	42	7.04	7.06	7.23	7.29
9	6.31	6.31	6.33	7.66	43	7.13	7.07	6.78	6.71
10	6.35	6.27	6.20	6.13	44	7.16	7.11	6.91	7.12
11	6.39	6.48	6.58	6.50	45	7.20	7.20	7.17	7.46
12	6.40	6.59	6.87	6.61	46	7.22	7.62	7.52	7.73
13	6.45	6.34	6.34	6.30	47	7.23	6.97	6.96	6.99
14	6.45	6.47	6.53	6.51	48	7.35	7.34	7.24	7.47
15	6.46	6.55	6.52	6.61	49	7.41	7.56	7.62	7.63
16	6.47	6.41	6.39	6.23	50	7.53	7.49	7.15	7.63
17	6.48	6.46	6.51	6.50	51	7.54	7.52	7.00	7.41
18	6.53	6.66	6.86	6.67	52	7.66	7.86	7.36	7.89
19	6.55	6.93	6.94	6.98	53	7.66	7.70	8.27	7.43
20	6.57	6.66	6.63	6.78	54	7.69	7.74	7.14	7.69
21	6.57	6.54	6.86	6.52	55	7.69	7.56	7.62	7.44
22	6.59	6.66	6.86	6.73	56	7.69	7.67	7.62	7.09
23	6.60	6.80	6.87	6.86	57	7.71	8.06	7.48	8.04
24	6.65	6.90	6.81	7.04	58	7.71	7.69	8.12	7.57
25	6.70	6.77	6.78	6.96	59	7.72	7.81	7.28	7.83
26	6.78	6.77	6.85	6.85	60	7.77	7.62	7.52	7.52
27	6.82	6.87	6.63	6.91	61	7.80	7.81	7.28	7.81
28	6.82	6.54	6.67	6.53	62	7.82	7.74	8.15	8.13
29	6.82	6.58	6.90	6.59	63	7.94	7.66	7.52	7.52
30	6.84	6.89	6.81	7.02	64	8.07	8.07	8.09	8.05
31	6.86	6.60	6.70	6.61	65	8.08	8.11	7.90	8.18
32	6.89	6.76	6.49	6.78	66	8.12	8.06	8.12	8.04
33	6.89	6.86	6.84	6.78	67	8.18	8.11	7.94	8.15
34	6.89	6.43	6.85	6.13	68	8.35	8.02	8.14	7.86

[a] The data were taken from ref 12. [b] The number of outliers is 21 (the definition of outlier is described in the footnote of Table V).

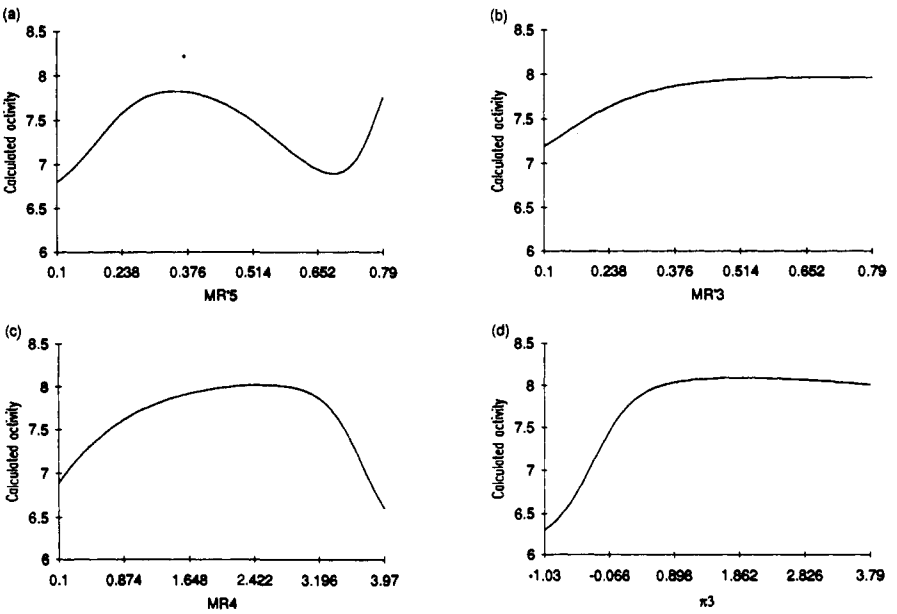

Figure 5. Biological activity as a function of the individual physicochemical parameter.

tivity and the predicted activity of any similar compound would differ by at most 0.62. In the light of this, one can be confident that the neural network is able to provide reliable predictions of biological activities of novel pyrimidine variants.

Dependence of Biological Activity on the Physicochemical Parameters. The relative dependence of each parameter in affecting the biological activities was investigated. It is anticipated that the identification of the functional dependences of biological activity to the physicochemical parameters could assist medicinal chemists.

The variation of the activity was monitored by changing the value of one input while keeping the remaining three inputs of the neural network constant at a quarter of their maximum ranges. On the basis of the resulting plots, the biological activity seems to be approximately a linear function of MR'_3 (Figure 5b) and it has nonlinear depen-

Table V. Comparison of Residual Variance, Rank Correlation Coefficient, and Number of Outliers for Neural Network and Regression Analysis

	residual variance	rank correlation	number of outliers[a]
neural network	0.029	0.94	12
regression analysis	0.075	0.88	25

[a] An outlier had $|\text{activity}_{obs} - \text{activity}_{pred}| > 0.25$.

dence upon the other three parameters. The plots suggest that there is a cubic dependence of MR'_5 (Figure 5a) and a parabolic dependence of MR_4 (Figure 5c). Moreover, the function shown in Figure 5d is different from the others, and it does not appear to be a simple mathematical function.

In Hansch's earlier analysis on this set of compounds,[12] a correlation equation was formulated:

$$\log (1/K) = 0.95MR'_5 + 0.89MR'_3 + 0.80MR_4 - 0.21MR_4^2 + 1.58\pi'_3 - 1.77 \log (\beta 10^{\pi 3} + 1) + 6.65 \quad (1)$$

where $\log \beta = 0.175$, $MR_4^0 = 1.85$, and $\pi_3^0 = 0.73$.

It was interesting to note that the neural network was consistent with some of their findings. Firstly, both the neural-network model and regression analysis agreed that the biological activity has a linear dependence of MR'_3 and a parabolic dependence of MR_4. Secondly, if one plots the correlation equation adopting π_3 as the only variable and keeping the other three parameters as constants, one discovers that the appearance of this plot is extremely similar to the corresponding one shown in Figure 5d. Thirdly, the optimum values MR_4^0 and π_3^0 also give very high activities in the corresponding neural network plots. Both of them would yield activities above 95% of their maximum values in their ranges of the plots.

The main discrepancy between the neural-network model and regression analysis seems to be the functional dependence of MR'_5. The regression equation fits the MR'_5 values with a simple linear term while the neural network predicts there may be a cubic relationship. An investigation was performed to show whether the inclusion of a cubic MR'_5 term would improve the quality of the QSAR. The following correlation equation for the inhibitory effects was obtained:

$$\log (1/K) = 11.79MR'^3_5 - 15.74MR'^2_5 + 6.55MR'_5 + 0.89MR'_3 + 0.80MR_4 - 0.21MR_4^2 + 1.58\pi'_3 - 1.77 \log (\beta 10^{\pi 3} + 1) + 6.24 \quad (2)$$

where $\log \beta = 0.175$.

For the purpose of comparison, this equation is constructed in a way such that each physicochemical parameter takes the same percentage of overall variance as they would in eq 1. Consistent with this is the fact that the congeners with either the maximum value (0.79) or the minimum value (0.10) of MR'_5 will be unaffected by the new equation. The calculated activities of the eight congeners which have activities different from those given by eq 1 are shown in Table VI. The improvement is noticeable. The RVs of the eight compounds are 0.074 and 0.093 for the cubic fit and the linear fit, respectively. This result further underlines the key strength of the neural network in performing this sort of implicit nonlinear mapping. It is deemed to be the main reason that neural network can outperform regression analysis in QSAR.

Table VI.[a] Comparison of the Calculated Activities from Regression Equations with a Cubic and Linear MR'_5 Term

number	observed activity	cubic MR'_5	linear MR'_5
7	6.28	6.66	6.79
9	6.31	6.18	6.33
15	6.46	6.70	6.52
16	6.47	6.56	6.39
30	6.84	6.99	6.81
39	6.97	7.35	7.44
42	7.04	7.08	7.23
43	7.13	6.77	6.78

[a] The other 60 compounds not listed in this table have identical calculated activities in both correlation equations.

Building a regression equation as complex as eq 2 cannot be inspired by a flash of brillance: it requires a laborious development phase. In regression analysis, the inclusion of nonlinear terms is on a trial and error basis. For neural networks this is not necessary. The researcher may simply consider the shapes of these plots and propose whether and how the nonlinear terms should be included in their analysis.

Conclusion

The results presented here add to the growing support for the use of neural networks in QSAR studies, but also emphasize just how much care needs to take in the design of the network. Not only are the results superior to regression analysis when judged in statistical terms but they also provide accurate predictions of activities of the compounds and furthermore permit the highlighting of the functional form those molecular parameters which play an important role in determining biological activity.

Acknowledgment. We thank David Livingston, David Manallack, and Colin Edge for their useful discussions. We are grateful to Vincent Lok, Steven Li, and Steve Chiu for their advice in the coding of the neural network simulator.