Development of Neural Network Simulator for Structure−Activity Correlation of Molecules (NECO). Prediction of Endo/Exo Substitution of Norbornane Derivatives and of Carcinogenic Activity of PAHs from ^{13}C-NMR Shifts

Yoshimi Isu,[†] Umpei Nagashima,[†] Tomoo Aoyama,[‡] and Haruo Hosoya*,[†]

Department of Information Sciences, Ochanomizu University, Bunkyo-ku, Tokyo 112, Japan, and Faculty of Engineering, Miyazaki University, Gakuenkihanadai, Miyazaki 889-21, Japan

Received September 6, 1995[⊗]

A perceptron type neural network simulator for structure−activity correlation of molecules has been developed with two different learning methods, i.e., back-propagation and reconstruction methods. First by use of the back-propagation method the exo/endo branching of norbornane and norbornene derivatives was correctly predicted from the set of ^{13}C NMR chemical shifts for various ring carbon atoms. Then the obtained correlation was analyzed by the reconstruction learning method. It was shown in this case that the NMR shifts for two carbon atoms out of seven have strong correlation with the exo/endo branching. Further, structure−activity correlation between the ^{13}C NMR chemical shifts and carcinogenicity of 11 polycyclic aromatic hydrocarbons was also analyzed using the reconstruction method. It was demonstrated that neural network analysis is suitable for the elucidation of complicated structure−activity problems where many factors are nonlinearly entangled.

INTRODUCTION

It has recently been recognized that application of neural network (NN) analysis offers great advantage over the conventional QSAR or QSPR method especially for the cases where the targeted properties are not linearly dependent on the chosen parameters.[1−10]

One of the present authors has developed the perceptron type NNs and shown their powerful utility for various nonlinear phenomena.[1−4,11,12] For performing systematic and efficient numerical analyses we have developed a NEural network simulator for structure−activity COrrelation of molecules, **Neco**, in C language for the portability and utility of high performance workstations.[13,14] In principle it can also be applied to many areas of problems even outside of science.

The NN is a data processing method which was derived from the neuron's operation in our brain. The main characteristics of the operations of a NN have been found to be nonlinear. Furthermore, the relationship between the input and output is automatically accumulated in the learning phase in the form of the weight matrix. Thus the NN method has widely been applied to data processing in which relationship between the input and output is difficult to be analyzed.

Chemists often ask if a given substance among a variety of compounds with similar structure has a specific property, useful or harmful. They also ask if two or more different properties of a given substance are really correlated, and if so how they are correlated. The QSAR studies have solved many of these two types of problems, and most of the conventional techniques have dealt with numerical problems.[15] Use of Graph-theoretical technique has accelerated the elucidation of QSAR problems.[16−21]

On the other hand, we have applied our neural network simulator, **Neco**, to two problems in which some specific property of a given substance, i.e., a set of ^{13}C-NMR chemical shifts of the component carbon atoms, has a crucial role in determining another specificity, such as molecular conformation or carcinogenicity.

One of the shortcomings of primitive NN analysis was that although good correlation between two or more seemingly irrelevant quantities is disclosed by the analysis it is usually difficult to trace the entangled network and to derive its causality. In our **Neco** by introducing the reconstruction method we could trace the thread of main correlation.[2−4,11,12,22,23]

A NN is composed of a number of connected neurons. Many types of network connections are known such as the Hopfield type,[24] in which all neurons are connected with each other, and the perceptron type with multilayer network connections. In this study we have adopted a perceptron type multilayer NN where the strength of the connection between neurons is preliminarily determined using well-established knowledge, which is called learning. The back-propagation method is usually applied for the learning of this type NN.

Details of the relationship between input and output are kept in the strength of the network connection by the learning. Since the information on the relationship between input and output is distributed in terms of the strength of the connection between neurons, the character of the relationship is difficult to be analyzed. Therefore analyses of the role of each input parameter to the specific output decision are usually not performed well. In order to overcome this problem, we have also used the weight matrix reconstruction method as a character analysis method. This method was first proposed by Aoyama and Ichikawa.[11]

In the following chapter the applied NN will briefly be explained. Analysis of the ^{13}C-NMR chemical shifts and the branching of norbornane and norbornene derivatives

[†] Ochanomizu University.
[‡] Miyazaki University.
[⊗] Abstract published in *Advance ACS Abstracts,* February 1, 1996.

0095-2338/96/1636-0286$12.00/0

using NN is demonstrated in the third chapter. In the fourth chapter the relationship between ^{13}C-NMR chemical shifts and the carcinogenicity of polycyclic aromatic hydrocarbons is discussed. The conclusion is given in the final chapter.

PERCEPTRON TYPE NEURAL NETWORK

1. The Applied Neural Network Structure. Though **Neco** can treat the neural networks consisting of several layers, we have applied a perceptron type neural network with three layers, i.e., input, hidden and output layers.[1,4,11,12] As shown in Figure 1, each layer consists of several neurons which are actually variables taking a value ranging from 0 to 1. One of the main characteristics of a perceptron type neural network is that the strength of the connection between neurons is preliminarily determined by the learning process, which will be briefly explained in the next section.

2. The Learning Method for Neural Network. 2.1. Back-Propagation Method. The relationship between the input and output data can automatically be analyzed in the learning phase of a perceptron type neural network.[2,4,11,12,22,23] The back-propagation method has been well established for the learning method of a perceptron type neural network. We adopted the back-propagation method for the learning procedure.

The procedure of the back-propagation method is as follows: The output value of neuron j in layer n, O_j, can be expressed by eq 1

$$O_j = f(y_j)$$

$$y_j = \sum_i W_{ij} x_i$$

$$f(x) = \frac{1}{1 + exp(-\alpha x)} \qquad (1)$$

where x_i is the output value of neuron i in layer $n - 1$ and $W_{i,j}$($= W_{i,j}^{(n,n-1)}$) is an element of the weight matrix and expresses the strength of the connection between the two neurons, i and j.

When an output vector, $\mathbf{O} = (O_1,O_2,...)$, is obtained for an input vector, this output vector **O** is compared with a vector which is called teaching vector, $\mathbf{t} = (t_1,t_2,...)$, and becomes a desirable output vector for the input. The learning here means to modify the weight value, $W_{i,j}$, so that the output vector **O** is nearly equal to the teaching vector **t** and the learning is carried out according to the following back-propagation algorithm.

Suppose that P sets of the input and teaching vectors are given. Let us define the evaluation function, E, for the learning as eq 2[22]

$$E = \sum_P \sum_j (O_j - t_j)^2 \qquad (2)$$

The learning of the given network is carried out until E becomes small enough according to the following equations

$$\delta W_{ij}^{(n-1,n)} = -d_j^{(n)} x_i \epsilon + \eta \delta W_{ij}^{(n-1,n)'} \qquad (3)$$

$$d_j^{(N)} = (O_j - t_j) f'(y_j) \qquad (4)$$

$$d_j^{(k)} = (\sum_l W_{jl}^{(k,k+1)} d_l^{(k+1)}) f'(y_j) \quad (1 \le k < N) \qquad (5)$$

where N is the number of layers. The second term of eq 3 avoids instability for steady convergence, $\delta W_{ij}^{(n-1,n)'}$ represents the correction in the last learning cycle, η is a parameter for stability of the learning, and ϵ is a parameter which determines a shift for correction. In this study, ϵ and η were tentatively set to be 0.75 and 0.8, respectively, to have a rapid convergence, whereas the results do not strongly depend on these parameters. Equation 4 is used only to correct the connection between the output and hidden layers, while eq 5 is for other connections. This procedure is applied to the knowledge, namely, the set of learning data. By correcting a set of learning data one cycle of learning is counted. Usually several thousands of cycles are required to attain convergence.

Since the network obtained by using the above back-propagation algorithm is usually very complicated, it is difficult to trace the entangled relationship between the input and output. Aoyama and Ichikawa suggested that the network can be simplified by introducing a forgetting process in the learning phase.[11] Then by tracing the simplified connections, one can understand the role of the input parameters to a specific output decision. Such a learning method was named "reconstruction learning" and applied here as a method for analyzing the character of the relationship between the input and output data.

2.2. Reconstruction Method. In a forgetting process the absolute values of the weight matrices are lessened according to eq 6

$$W_{ij} = W_{ij} - \mathrm{sgn}(W_{ij})\zeta\{1 - \delta(W_{ij})\} \qquad (6)$$

$$\delta(W_{ij}) = \begin{cases} 0 & \text{if} |W_{ij}| > \zeta \\ 1 & \text{otherwise} \end{cases}$$

where ζ is a parameter which suppresses a rapid change of W_{ij} and is set to be about a tenth of ϵ.[11] As following the suggestion by Aoyama and Ichikawa, the value of ϵ was tentatively set to be 0.075 (one-tenth of the learning process) in this study so that E converges very slowly, because the learning must be repeated well enough to obtain a simplified network. The term in braces of eq 6 is appended so as not to change the sign of W_{ij} for the cases where the absolute value of W_{ij} is very close to 0.

After repeating the learning and forgetting processes, the information distributed among the connections between several neurons is put together. The training is stopped when the maximum change of weight becomes less than 1.0×10^{-4}. As a result, some connections become stronger and other unnecessary connections diminishingly small. As mentioned previously, the network obtained according to the above back-propagation algorithm is so complicated that it is rather difficult to trace correctly the entangled relationship between the input and output data. However, as shown in Figure 2, only those connections which are responsible for the correlation between the input and output survive in the reconstructed weight matrix, since the strengths of such connections among neurons are enhanced through the reconstruction iteration.

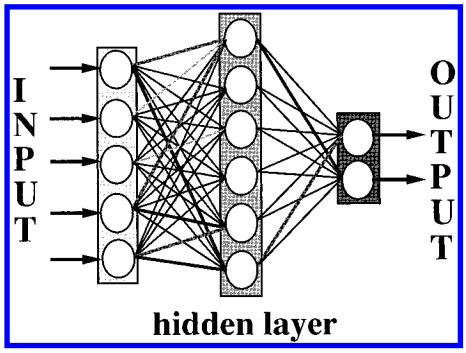

Figure 1. An example of a three-layered neural network, *N*(5,6,2). A circle represents a neuron which is connected to all the neurons in the neighboring layers. The value of weight for each connection can automatically be calculated by the back-propagation method.

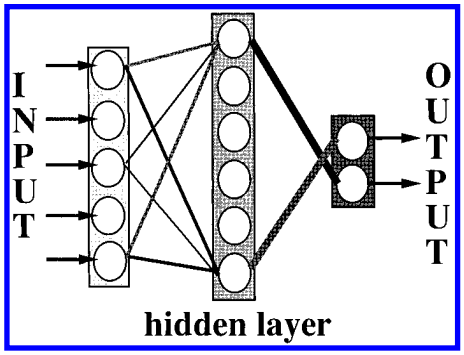

Figure 2. The network obtained by the reconstruction method. After repeating the learning and forgetting processes for the neural network given in Figure 1 the simplified network can be derived.

RELATIONSHIP BETWEEN ^{13}C-NMR CHEMICAL SHIFTS AND EXO/ENDO BRANCHING OF NORBORNANES

A number of NN analyses have been performed for predicting the NMR chemical shifts from the topological structure of molecules.[25−29] Contrary to these studies we have undertaken to predict several molecular properties by NN analysis using ^{13}C-NMR chemical shifts of a series of compounds. Namely, to test the predictability of **Neco**, the exo or endo branching of norbornane and norbornene derivatives was predicted from the set of the ^{13}C-NMR chemical shifts of ring carbon atoms by using the same data[30] which were chosen by Sasaki et al. in their learning machine and cluster analysis.[31] Our preliminary calculations have been given elsewhere.[1,11,12]

The numberings of carbon atoms of the norbornane skeleton are given in the bottom right of Figure 3. The structures, ^{13}C-NMR data, and exo/endo branchings of 38 compounds are given in Figure 3 and Table 1. The upper 25 (nos. 1−25) out of 38 data were used for the learning, and the rest of the 13 compounds were used for the prediction. The neural network structure used in this problem is the perceptron type network with three layers. The numbers of neurons in the input, hidden, and output layers are respectively 7, 14, and 1. This NN is denoted by *N*(7,14,1).[11]

For the learning process the set of ^{13}C shifts of 25 compounds were chosen as the input data, all of which were scaled between 0.1 and 0.9. The output data for exo and endo isomers were set to be 0 and 1, respectively. First, the learning was carried out by the back-propagation method until the evaluation function (eq 2) becomes 1.0×10^{-4} or

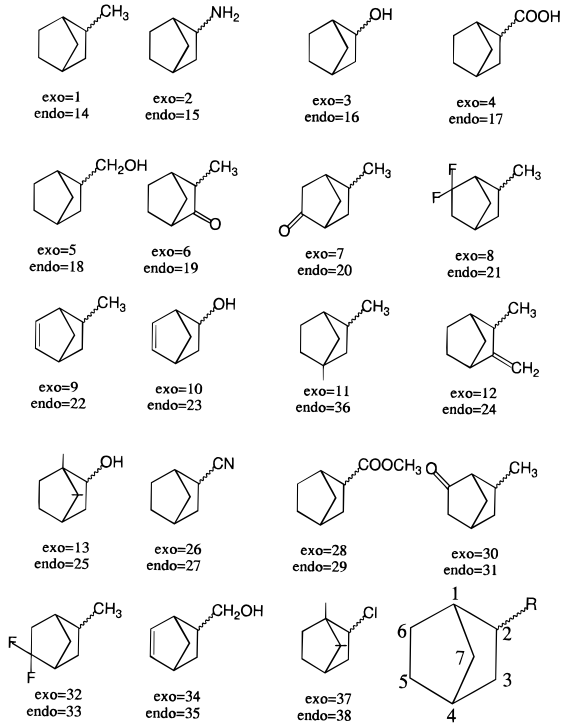

Figure 3. The structural formulas of norbornanes and norbornenes together with the numbering of the carbon atoms (bottom right). The number below each norbornane skeleton indicates the data number in Table 1.

less. After the learning the set of ^{13}C chemical shifts for 13 compounds was input to see if the correct prediction of exo or endo can be derived as the output. The predicted results are shown in Table 2. For 12 compounds out of 13 the isomer predictions were correctly obtained, whereas for compound no. 31 the prediction was wrong. The program previously written by Aoyama et al.[1,11,12] also could not correctly predict the exo/endo branching for the same compound. Sasaki et al. also studied the same problem with the linear learning machine and the cluster analysys.[31] However they could not correctly predict the exo/endo branching for two compounds, nos. 31 and 38. Our neural network method is shown to be better than the linear method.

In order to analyze the character of the relationship between the ^{13}C chemical shifts and the exo/endo branching, the reconstruction learning method was used. The weight matrix between the input and hidden layers is shown in Table 3, which was obtained after repeating the reconstruction learning for 3.0×10^5 times. As is evident from Table 3, the reconstructed weight matrices are very simplified. The neurons 6 and 7 (column) in the input layer are shown to have strongest connections with neurons in the hidden layer. Since the number of neurons in the input layer corresponds to the number of the carbon atoms, it is found that the chemical shifts for C_6 and C_7 have best correlation with the exo/endo branching. Following C_6 and C_7, the ^{13}C shifts for C_5, C_2, and C_4 are found to have fair correlation with the exo/endo branching.

The importance of the chemical shifts for C_6 and C_7 was suggested by the reconstruction method. Namely to examine the reliability of this suggestion, the learning of the relationship between the ^{13}C shifts only for C_6 and C_7 and the exo/endo branching was carried out by the back-propagation method. This time a NN of *N*(2,4,1) was used for the

Table 1. Relative ^{13}C-NMR Chemical Shifts and Exo/Endo Branching in Norbornanes[29]

no.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	exo/endo
				For Learning				
1	6.7	6.7	10.1	0.5	0.2	−1.1	−3.7	exo
2	8.9	25.3	12.4	−0.4	−1.2	−3.1	−4.4	exo
3	7.7	44.3	12.3	−1.0	−1.3	−5.2	−4.4	exo
4	4.6	16.7	4.4	−0.2	−0.3	−1.0	−1.8	exo
5	1.8	15.1	4.4	−0.2	0.2	−0.7	−3.3	exo
6	5.7	3.0	2.6	−0.5	−0.4	0.7	−3.5	exo
7	6.1	5.9	10.6	0.6	0.2	0.2	−3.7	exo
8	6.5	6.3	10.4	0.3	−0.8	−0.1	−3.5	exo
9	6.5	7.5	9.5	0.5	1.7	0.7	−3.8	exo
10	7.8	47.0	11.7	−1.3	3.9	−2.7	−3.2	exo
11	6.9	6.9	10.1	0.7	−1.2	0.1	−3.9	exo
12	5.6	4.9	7.0	0.2	−1.1	0.2	−3.9	exo
13	2.5	42.5	11.9	−0.8	−1.1	−2.4	1.4	exo
14	5.4	4.5	10.6	1.4	0.5	−7.7	0.2	endo
15	6.8	23.3	10.5	1.2	0.6	−9.5	0.3	endo
16	6.3	42.4	9.5	0.9	0.2	−9.7	−0.9	endo
17	4.2	16.2	2.1	0.9	−0.6	−4.8	1.9	endo
18	1.7	12.8	4.0	0.4	0.2	−7.2	1.4	endo
19	4.7	3.1	2.2	0.3	1.3	−6.5	−0.6	endo
20	4.7	5.3	9.2	1.3	−0.4	−6.5	1.4	endo
21	4.6	11.5	8.9	−0.1	0.8	0.4	1.8	endo
22	5.6	7.5	8.7	1.4	1.7	−3.0	1.7	endo
23	7.1	47.8	13.3	2.2	3.6	−3.4	0.3	endo
24	4.1	4.2	7.0	0.7	0.5	−7.4	0.0	endo
25	3.2	40.2	10.4	−0.5	0.0	−10.3	3.1	endo
				For Prediction				
26	5.5	1.0	6.3	−0.3	−1.5	−1.6	−1.3	exo
27	3.4	0.1	5.5	0.2	−0.7	−4.9	0.0	endo
28	5.1	16.4	4.2	−0.4	−1.1	−1.4	−2.1	exo
29	4.0	15.9	2.2	0.7	−0.7	−5.0	1.7	endo
30	6.6	7.0	10.1	0.2	−1.2	0.5	−3.7	exo
31	6.0	8.4	11.2	−0.1	0.7	−1.5	−1.6	endo
32	6.3	7.2	9.8	0.7	−0.1	0.8	−3.5	exo
33	5.1	4.8	8.4	1.1	−0.1	−7.3	1.6	endo
34	1.9	17.1	5.2	−0.1	0.9	0.9	−3.4	exo
35	2.3	18.3	5.0	0.3	1.3	−2.9	1.4	endo
36	5.1	4.0	8.4	1.1	0.2	−7.7	1.6	endo
37	2.9	30.3	13.4	−0.5	−2.1	−0.7	2.0	exo
38	3.7	29.8	10.8	−1.6	−1.1	−9.0	2.2	endo

Table 2. Untrained Output Data and Comparison between the Predicted and Observed Branching

	output		exo/endo	
no.	C_1-C_7[a]	C_6, C_7[b]	predicted[c]	obsd
26	0.1468	0.0001	exo	exo
27	0.9653	1.0000	endo	endo
28	0.0280	0.0000	exo	exo
29	0.9899	1.0000	endo	endo
30	0.0049	0.0001	exo	exo
31	0.2332	0.0000	exo	endo
32	0.0169	0.0001	exo	exo
33	0.9958	1.0000	endo	endo
34	0.0131	0.0001	exo	exo
35	0.9828	0.9277	endo	endo
36	0.9963	1.0000	endo	endo
37	0.1390	0.3420	exo	exo
38	0.9456	1.0000	endo	endo

[a] C_1-C_7 chemical shifts were used for the calculation. [b] Only C_6 and C_7 shifts were used. [c] Predicted forms are the same for both the cases.

analysis. The set of data of the same size were used for the learning, i.e., 25 (nos. 1−25) out of 38 data (Table 1). Table 2 shows the predicted exo/endo branchings of the untrained 13 compounds. The result completely agrees with what was obtained by using ^{13}C shifts for seven carbon atoms. This suggests that the exo/endo branching in this case can be predicted only by using ^{13}C shifts for C_6 and C_7.

Table 3. Reconstructed Weight Matrix, $W^{1,2}_{i,j}$, between the First (Column) and the Second (Row) Layers

j/*i*	1	2	3	4	5	6	7
1	0.00	0.01	0.00	0.00	0.00	0.01	0.00
2	0.00	0.01	0.01	0.00	0.00	0.01	0.00
3	0.00	0.01	0.00	0.00	0.00	0.00	0.00
4	0.00	0.01	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.01	0.00
6	0.01	0.00	0.00	0.01	0.00	−0.00	0.00
7	0.00	0.01	0.01	0.01	0.00	−0.00	**2.07**
8	0.00	0.00	0.00	0.01	0.00	0.00	0.00
9	−0.01	**2.93**	0.00	**−2.27**	**−3.28**	**4.32**	**−4.82**
10	0.01	0.00	0.00	0.01	0.01	0.00	0.00
11	0.00	0.01	0.01	0.00	0.01	0.00	0.01
12	0.00	0.00	0.01	0.00	0.01	0.01	0.00
13	0.00	0.00	0.00	0.01	0.00	0.01	0.00
14	0.00	0.01	0.01	0.01	0.00	0.00	0.01

Then let us try to draw a correlation map which can predict the exo/endo branching just from the combination of the values of C_6 and C_7. By using the back-propagation method for $N(2,4,1)$ one can calculate the output value for any combination of C_6 and C_7 values. Figure 4a shows such a boundary curve that separates the exo and endo regions from each other. This curve was obtained by continuously plotting those points which give the output value of 0.5. The ranges of the values of C_6 and C_7 are respectively chosen to be −10.3 to 0.7 and −4.4 to 3.1 for the groups of the 38 compounds. Note that the obtained boundary is not linear. Further on this map all 38 points are plotted.

Note that except for no. 31 all the studied and predicted points of exo and endo isomers lie respectively in the exo and endo regions. If all other NMR data are correct, the relative ^{13}C shift for C_6 of no. 31 compound should be much smaller than the reported one. On the other hand, if the no. 31 data is included in the learning, the boundary curve is obtained to be as in Figure 4b. In this case the no. 31 point is correctly bound in the endo region, while the point predicted for no. 26 compound steps out into the wrong region. From these experiences one can say that in order to get good prediction from NN analysis the data for learning process should have high quality.

Although the ^{13}C shifts for C_6 and C_7 are crucial for predicting the exo/endo branching, one can see from the reconstructed weight matrix (Table 3) that the chemical shift for C_5 also shows fair correlation with the branching. Thus the learning of the relationship among these three chemical shifts and the exo/endo branching was carried out by the back-propagation method. The result is shown in Figure 5. The meshed curve surface was drawn for the region with 0.5 output value. The region of exo is above the surface, while endo is below the surface. The border surface shown in Figure 5 is not flat but curved. This nonlinearity is the main characteristic of the neural network analysis.

Instead of C_5 the importance of the ^{13}C shift for C_4 was pointed out by Sasaki et al.[31] Therefore the learning of the relationship between the chemical shifts for C_4, C_6, and C_7 and the exo/endo branching was also carried out by the back-propagation method. The obtained border surface separating exo and endo isomer is shown in Figure 6. The region of exo is above the surface, while endo is below the surface. As seen from Figure 6, the border surface is not curved but flat. This means that the linear method can yield only linearly described results, whereas our neural network method can analyze well nonlinear relationships.

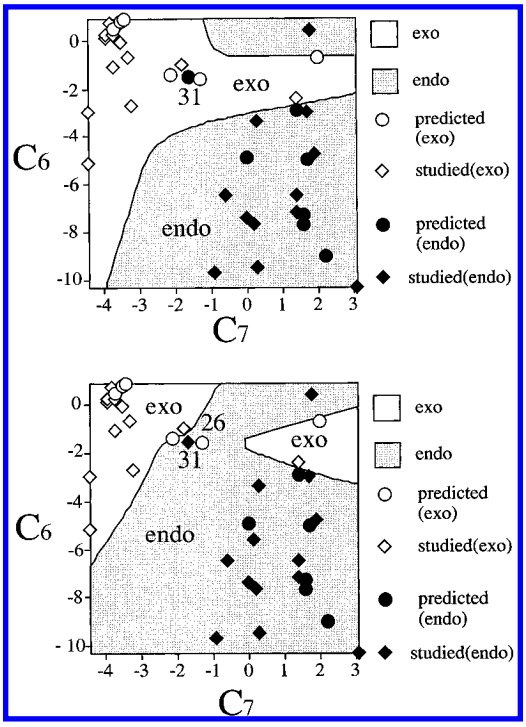

Figure 4. Correlation map of exo/endo branching predicted from the two chemical shifts for C_6 and C_7. (a) The no. 31 data was included for the learning. (b) The no. 31 data was excluded from the learning.

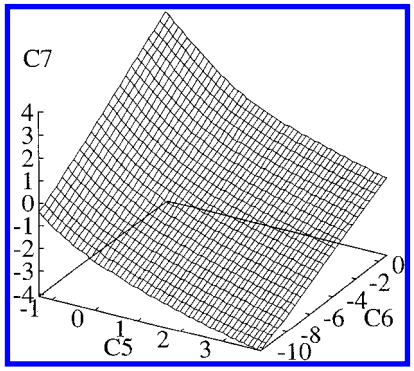

Figure 5. The border surface separating the exo (above) and endo (below) regions described by the set of three parameters, C_5-C_7.

Finally we should mention some chemistry. By observing a stereo model (see Figure 7) it is easily anticipated that C_6 and C_7 are the most sensitive to the exo/endo branching, followed by the farthest C_5, since all these carbon atoms lie off the plane determined by C_1-C_4, which are insensitive to the direction of the exo/endo branching.

CARCINOGENIC ACTIVITY OF POLYCYCLIC AROMATIC HYDROCARBONS

The relationship between the structure and the carcinogenic activity seems very complicated and nonlinear. Various theoretical analyses have been performed for the origin of the carcinogenicity of polycyclic aromatic hydrocarbons (PAH). In 1939 Schmidt reported that the carcinogenicity of PAH has a close relationship with the distribution of π-electrons.[32] According to his analysis, PAHs with large π-electron density in L-region (Figure 8) have strong carcinogenic activity. Pullman and Pullman proposed a hypothesis that coupling of high π-electron density in K-region and low density in L-region induces strong carcinogenic activity to PAH.[33,34] Importance of the bay-region and also of the metabolic transformation of carcinogenicity of PAH has been pointed out by Jerina et al.[35] They suggested that diol epoxide formed at the bay-region is more responsible for the carcinogenic activity of PAHs than other diol epoxides. In these studies quantum mechanically calculated quantities were mainly used as indices to predict the carcinogenic activity. Several attempts have also been made for analyzing this carcinogenic problem by NN techniques.[36,37]

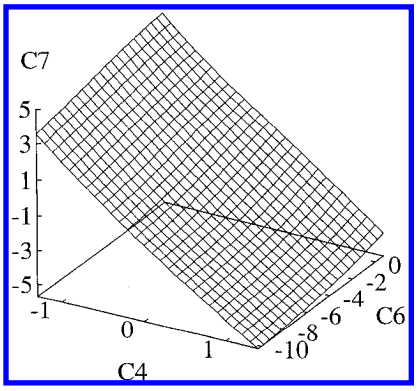

Figure 6. The border surface separating the exo (above) and endo (below) regions described by the set of three parameters, C_4, C_6, and C_7.

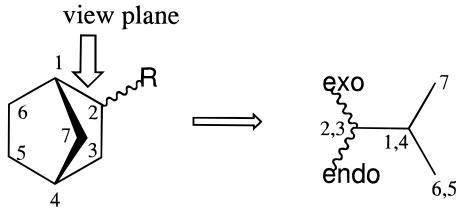

Figure 7. Structure of norbornane skeleton viewed from the axis passing through the plane determined by C_1-C_4.

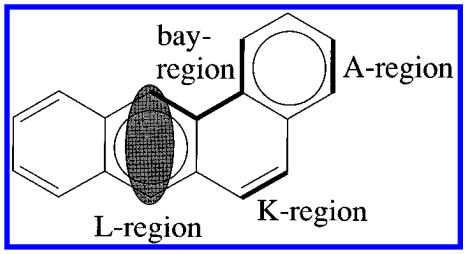

Figure 8. The K-, L-, A-, and bay-regions of a polycyclic aromatic hydrocarbon.

We have applied **Neco** for analyzing the relationship between the experimentally derived ^{13}C shifts and the carcinogenic activity of PAH.[38] The structural formulas for 11 PAHs used in this study are depicted in Figure 9 together with the numbering of carbon atoms. Note that the geometrical structure of the path from C_1 to C_{10} is the largest common skeletal carbon moiety of the chosen 11 PAHs, although only compound no. 6 contains a sulfur atom. We could collect a set of ^{13}C-NMR chemical shifts for those ten carbons of all the PAH molecules with and without carcinogenicity.[39,40] The set of these ten ^{13}C shifts for the 11 PAHs are shown in Table 4. The numbers of carbon atoms and compounds in Table 4 are the same as those in Figure 9. In this case all the data of 11 PAHs were used for the learning process.

The learning of the relationship between the ^{13}C shifts and the carcinogenicity was carried out by repeating 3.0×10^5 times the reconstruction learning method. The $N(10,10,1)$

Table 4. The ^{13}C-NMR Chemical Shifts for Ten Carbons from Each of 11 PAHs and Their Carcinogenic Activities

no.	activity	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	ref
1	+	127.1	132.0	128.6	126.8	126.8	122.9	130.7	130.6	121.5	128.9	40
2	++	127.2	132.1	128.6	126.8	127.0	122.9	130.3	130.8	122.4	129.2	39
3	+++	124.7	131.3	128.9	125.9	125.9	122.9	128.3	127.4	122.1	127.4	39
4	+	126.9	131.7	128.3	126.8	126.8	123.6	130.5	130.9	120.3	128.7	40
5	+++	125.5	132.2	127.4	126.4	126.3	130.2	130.5	131.3	127.7	129.1	40
6	+	126.3	129.9	120.7	126.8	119.1	139.1	131.6	131.6	139.1	119.1	39
7	+	129.5	130.5	123.9	127.6	127.6	123.9	130.5	129.5	120.5	126.4	39
8	−	129.7	129.7	123.2	127.1	127.1	123.2	129.7	129.7	123.2	127.1	40
9	−	128.5	130.1	123.4	127.5	127.6	123.7	130.1	130.2	123.4	127.5	39
10	−	126.4	133.7	127.4	126.3	131.6	136.0	130.2	131.4	127.2	128.0	39
11	−	127.5	132.6	128.6	126.7	126.3	123.2	130.6	128.3	121.3	126.7	39

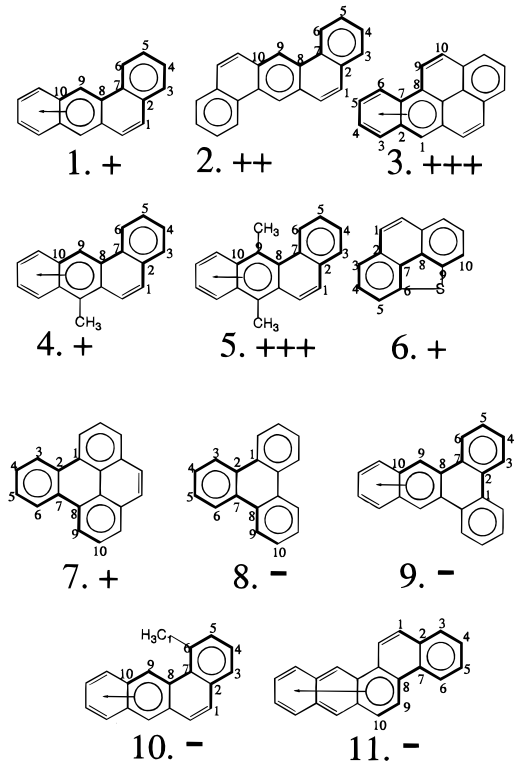

Figure 9. The structural formulas and numbering of carbon atoms of PAHs used in our analysis. (−) noncarcinogenic, (+, ++, +++) carcinogenic activity in increasing order. 1: benz[*a*]anthracene,[34,35,41] 2: dibenz[*a*,*h*]anthracene,[34,43] 3: benzo[*a*]pyrene,[34,35,43] 4: 7-methylbenz[*a*]anthracene,[44] 5: 7,12-dimethylbenz[*a*]anthracene,[44] 6: phenanthro[4,5-*b*,*c*,*d*]thiophene,[40] 7: benzo[*e*]pyrene,[42] 8: triphenylene,[34,43] 9: dibenz[*a*,*c*]anthracene,[34,41,43] 10: 1-methylbenz[*a*]anthracene,[42] 11: benzo[*b*]chrysene.[41]

Table 5. Reconstructed Weight Matrix between the Input (Column) and Hidden (Row) Layers

	1	2	3	4	5	6	7	8	9	10
1	0.00	0.01	0.01	−0.00	0.01	0.01	0.01	−0.00	**4.28**	0.01
5	**−7.21**	−0.01	**2.75**	−0.01	−1.86	−3.52	−0.01	**4.62**	−0.01	0.01
9	3.42	0.00	−0.01	**5.27**	−0.01	0.00	0.01	−0.00	−1.79	**−10.18**

network structure was taken as an initial NN structure because the number of the units for hidden layers are optimized by the reconstruction method. The reconstructed weight matrix between the input and hidden layers is listed in Table 5. Optimized NN structure was *N*(10,3,1).

The number of columns in Table 5 is the number of neurons in the input layer and corresponds to the number of carbon atoms giving ^{13}C shifts shown in Figure 9. The number of rows in Table 5 is the number of neurons in the hidden layer. Those neurons which have no connection with the neurons in the input layer are omitted from Table 5.

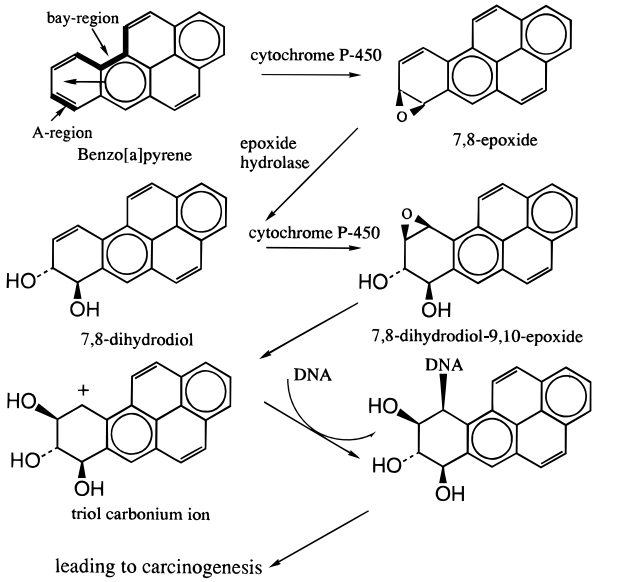

Figure 10. Mechanism of metabolic transformation leading to carcinogenesis by benzo[*a*]pyrene.[41]

The neurons of C_{10} and C_1 were found to have strong connections with the neurons in the hidden layer. Further the neurons of C_4, C_8, and C_9 were found to have also fair connections with the neurons in the hidden layer. This suggests that the ^{13}C shifts for C_{10} and C_1 have strong correlation with carcinogenicity, while C_4, C_8, and C_9 also have fair correlation. We can expect that C_9 and C_{10} atoms correspond to the carbons which are located in or near the L-region, while C_1 is in the K-region, and C_8 and C_9 are in the bay-region. The C_4 carbon is located in the region which is called the A-region,[41] which is the site of the first epoxidation in metabolic transformation leading to carcinogenesis (Figure 10).

The learning of the relationship between the carcinogenicity and the ^{13}C shifts only for C_{10} and C_1 was carried out by the back-propagation method. After learning that just by drawing the border line with 0.5 output value as in Figure 4, we could obtain the classification map separating the carcinogenic and noncarcinogenic regions as Figure 11a. In this analysis we have intentionally included the data of such a compound as no. 6 which does not contain the bay-region. As evident from Figure 11a the ^{13}C shifts for C_{10} of noncarcinogenic compounds range from about 126−128 ppm. On the other hand, Figure 11b was obtained by excluding the data of compound no. 6 for the learning. The resultant classification is almost the same as the case where no. 6 was included for the learning. Again we could draw the border surface separating the carcinogenic and noncarcinogenic compounds by the values of ^{13}C shifts for C_{10},

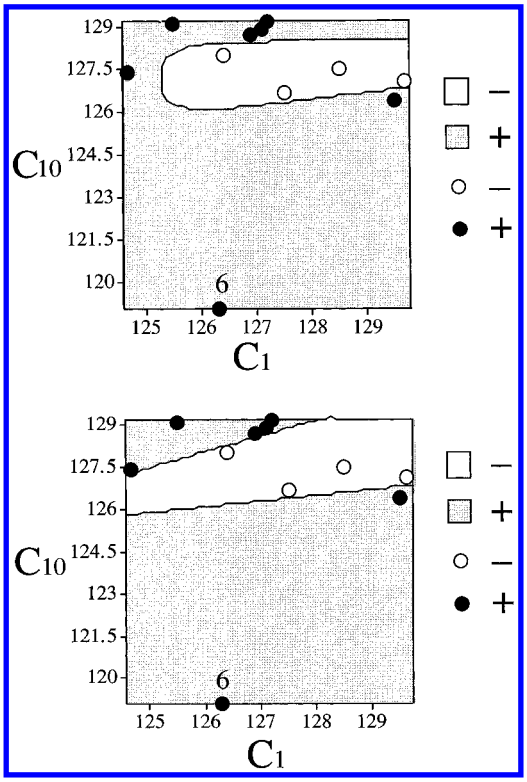

Figure 11. Classification map separating the carcinogenic (+) and noncarcinogenic (−) groups from the ^{13}C-NMR chemical shifts (in ppm) for C_{10} and C_1. (a) Compound no. 6 was included for the learning. (b) Compound no. 6 was excluded from the learning.

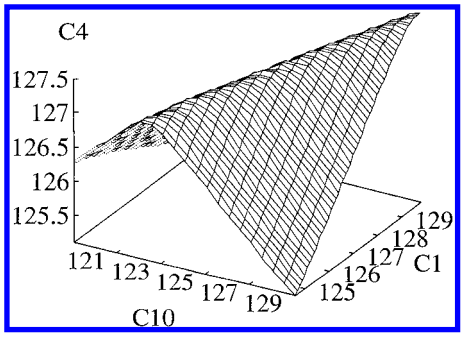

Figure 12. Classification map separating the carcinogenic (above) and noncarcinogenic (below) groups from the ^{13}C-NMR chemical shifts (in ppm) for C_{10}, C_1, and C_4.

C_1, and C_4 (Figure 12). The carcinogenic compounds are found to be located above the meshed surface drawn in Figure 12, while the noncarcinogenics are located below the surface. As seen from Figure 12, the carcinogenicity has strong nonlinear correlation with the ^{13}C shifts for C_{10}, C_1, and C_4 atoms.

In the analyses by NNs, the nonlinear border surface such as shown in Figure 12 can automatically be obtained revealing that a nonlinear mechanism is involved. These findings suggest that a NN analysis is suitable for data processing in which linear analysis may fail to be applied.

CONCLUDING REMARKS

A perceptron type NEural network simulator for structure−activity COrrelation of molecules, **Neco**, has been developed with two different learning methods, i.e., back-propagation and reconstruction methods.

Neco was applied to the prediction of exo/endo isomers of norbornane and norbornene derivatives from ^{13}C NMR chemical sifts. The exo/endo isomers of 12 out of 13 norbornane derivatives could correctly be predicted from the set of observed chemical shifts, after studying the correlation among the set of chemical shifts and exo/endo forms for 25 derivatives. This result is shown to be superior to that obtained by a learning machine and cluster analysis. The observed correlation was analyzed by the reconstruction learning method, revealing that only two ring carbons, C_6 and C_7, out of seven have strong correlation with the exo/endo forms. This result can be well explained by the stereochemical consideration.

The relationship between the ^{13}C-NMR chemical shifts and the carcinogenicity of 11 polycyclic aromatic hydrocarbons was also analyzed by the reconstruction learning method. It was suggested that the ^{13}C shifts for the five carbons in L-, K-, A-, and bay-regions are strongly correlated with the carcinogenicity.

From our experience of the application of neural network, it was found that neural network analyses are suitable for data processing in which correlation among many factors is nonlinearly entangled.

ACKNOWLEDGMENT

The authors express their sincere thanks to Dr. Shigeru Ohshima, Dr. Yohko Sakamoto, and Mr. Toshikazu Aoki of Toho University for their data collection of carcinogenic PAHs and valuable advice.

REFERENCES AND NOTES

(1) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Pharmaceutical Problems. I. Method and Application to Decision Making. *Chem. Pharm. Bull.* **1989**, *37*, 2558−2560.
(2) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.
(3) Aoyama, T.; Ichikawa, H. Basic Operating of Characteristics of Neural Networks When Applied to Structure-Activity Studies. *Chem. Pharm. Bull.* **1991**, *39*, 358−366.
(4) Aoyama, T.; Ichikawa, H. Obtaining the Correlation Indices between Drug Activity and Structural Parameters Using a Neural Network. *Chem. Pharm. Bull.* **1991**, *39*, 372−378.
(5) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley: New York, 1990.
(6) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Anal. Chim. Acta* **1991**, *248*, 1−30.
(7) 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.
(8) Maggiora, G. M.; Elrod, D. W.; Trenary, R. G. Computational Neural Networks as Model-Free mapping Devices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 732−741.
(9) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.
(10) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Publishers: New York, 1993.
(11) Aoyama, T.; Ichikawa, H. Reconstruction of Weight Matrices in Neural Networks−A Method of Correlating Outputs with Inputs. *Chem. Pharm. Bull.* **1991**, *39*, 1222−1228.
(12) Aoyama, T.; Ichikawa, H. Neural Networks as Nonlinear Structure−Activity Relationship Analyzers. Useful Functions of the Partial Derivative Method in Multilayer Neural Networks. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 492−500.
(13) Isu, Y.; Nagashima, U.; Hosoya, H.; Aoyama, T. Development of NEural Network Simulator for Structure−Activity COrrelation of Molecules: Neco (in Japanese). *J. Chem. Software* **1994**, *2*, 76−95.
(14) Isu, Y.; Nagashima, U.; Hosoya, H.; Ohshima, S.; Sakamoto, Y.; Aoyama, T. Development of NEural Network Simulator for Structure−Activity COrrelation of Molecules: Neco (2) (in Japanese). *J. Chem. Software* **1996**, in press.
(15) Hansch, C.; Fujita, T. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(16) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(17) Platt, J. R. Prediction of Isomeric Differences in Paraffin Properties. *J. Phys. Chem.* **1952**, *56*, 328−336.
(18) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(19) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological Indices for Structure−Activity Correlations. *Top. Curr. Chem.* **1983**, *114*, 21−55.
(20) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.
(21) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modeling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lipais, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984; p 745.
(22) McClelland, J. L.; Rumelhart, D. E. *Parallel Distributed Processing*; McClelland, J. L., Rumelhart, D. E., PDP Research Group, Eds.; MIT Press: Cambridge, MA, 1988; Vol. 1.
(23) Minsky, M.; Papert, S. *Perceptron−An Essay in Computational Geometry*; MIT Press: Cambridge, MA, 1969.
(24) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Capabilities. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 2554−2558.
(25) Meyer, B.; Hansen, T.; Nute, D.; Albersheim, P.; Darvill, A.; York, W.; Sellers, J. Identification of the ^{1}H-NMR Spectra of Complex Oligosaccharides with Artificial Neural Networks. *Science* **1991**, *251*, 542−544.
(26) Kvasnicka, V. An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63−76.
(27) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157−1164.
(28) Doucet, J. P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. Neural Networks and ^{13}C NMR Shift Prediction. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320−324.
(29) West, G. M. J. Predicting Phosphorus NMR Shifts Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 577−589.
(30) Kowalsky, B. R. *Chemometrics: Theory and Applications*; ACS Symposium Series 53; American Chemical Society: Washington, DC, 1977; p 423.
(31) Sasaki, S.; Abe, H.; Takahashi, Y.; Takayama, T.; Miyashita, Y. *Introduction to Pattern Recognition for Chemists* (in Japanese); Tokyo Kagaku Dojin: Tokyo, 1984; Chapter 1.
(32) Schmidt, O. Z. Density Distribution of B-electron. II. Characterization of Simple and of Cancer Producing Hydrocarbons through the Density Distribution of Certain Valence Electrons (B-electrons). *Physik. Chem.* **1939**, *B42*, 83−110.
(33) Pullman, A.; Pullman, B. *Rev. Sci.* **1946**, *84*, 145−158.
(34) Pullman, A.; Pullman, B. Electronic Structure and Carcinogenic Activity of Aromatic Molecules. *Ad. Cancer Res.* **1955**, *3*, 117−169.
(35) Jerina, D. M.; Lehr, R. E.; Schaefer, R. M.; Yagi, H.; Karle, J. M.; Thakker, D. R.; Wood, A. W.; Conney, A. H. Bay Region Epoxides of Dihydrodiols. A Concept Explaining the Mutagenic and Carcinogenic Activity of Benzo(a)pyrene and Benzo(a)anthracene. In *Origins of Human Cancer*; Hiatt, H., Watson, J. D., Winstin, I., Eds.; Cold Spring Harbor: New York, 1977; p 639.
(36) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.
(37) Villemin, D.; Cherqaoui, D.; Mesbah, A. Predicting Carcinogenicity of Polycyclic Aromatic Hydrocarbons from Back-Propagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1288−1293.
(38) Sakamoto, Y.; Watanabe, S. On the Relationship between the Chemical Structure and the Carcinogenicity of Polycyclic and Chlorinated Monocyclic Aromatic Compounds as Studied by Means of ^{13}C NMR. *Bull. Chem. Soc. Jpn.* **1986**, *59*, 3033−3038.
(39) Karcher, W.; Fordham, R. J.; Dubois, J. J.; Glaude, P. G. J. M.; Ligthart, J. A. M. *Spectral Atlas of Polycyclic Aromatic Compounds*; D. Reidel Publ.: Dordrecht, 1983.
(40) Ozubko, R. S.; Buchanan, G. W.; Smith, I. C. P. Carbon-13 Nuclear Magnetic Resonance Spectra of Carcinogenic Polynuclear Hydrocarbons. I. 3-Methylcholanthrene and Related Benzanthracenes. *Can. J. Chem.* **1974**, *52*, 2493−2501.
(41) Smith, I. A.; Berger, G. D.; Seybold, P. G.; Serve, M. P. Relationships between Carcinogenicity and Theoretical Reactivity Indices in Polycyclic Aromatic Hydrocarbons. *Cancer Res.* **1978**, *38*, 2968−2977.
(42) Cavalieri, E. L.; Rogan, E. G.; Roth, R. W.; Saugier, R. K.; Hakam, A. The Relationship between Ionization Potential and Horseradish Peroxidase/Hydrogen Peroxide−Catalyzed Binding of Aromatic Hydrocarbons to DNA. *Chem. Biol. Interact.* **1983**, *47*, 87−109.
(43) Allison, A. C.; Nash, T. Electron Donation and Acceptance by Carcinogenic Compounds. *Nature* **1963**, *197*, 758−763.
(44) Caspar, M. L.; Stothers, J. B.; Wilson, N. K. ^{13}C Nuclear Magnetic Resonance Studies of Methylated Anthracenes. *Can. J. Chem.* **1975**, *53*, 1958−1969.

CI950108B